For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
Universitatis
Albertensis

QUAECUMQUE VERA

Digitized by the Internet Archive
in 2021 with funding from
University of Alberta Libraries

https://archive.org/details/Treasure1975

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR	MORRIS RALPH TREASURE
TITLE OF THESIS	THE INFERRING AND HYPOTHESIZING ABILITIES OF JUNIOR HIGH SCHOOL STUDENTS
DEGREE FOR WHICH THESIS WAS PRESENTED	DOCTOR OF PHILOSOPHY
YEAR THIS DEGREE GRANTED	1975

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

THE UNIVERSITY OF ALBERTA

THE INFERRING AND HYPOTHESIZING ABILITIES

OF

JUNIOR HIGH SCHOOL STUDENTS

by

MORRIS RALPH TREASURE ©

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF SECONDARY EDUCATION

EDMONTON, ALBERTA

FALL, 1975

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, the thesis entitled, "The Inferring and Hypothesizing Abilities of Junior High School Students", submitted by Morris R. Treasure in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

A test of inferring ability and a test of hypothesizing was developed to gather information about the ability of junior high school students to exhibit these skills. These instruments were first validated and then used to test a series of eight hypotheses posed about the inferring and hypothesizing behavior of students in age groups of 11 to 15 on the battery of tests.

The review of the literature established the rationale for the study in the work of Piaget, Bruner and Gagné in their descriptions of cognitive growth, learning cycles and the hierarchical relationship among scientific processes. It was established that making inferences was a concrete operational skill, a skill learned relatively early in one's learning career and a skill learned before one could formulate hypotheses in the scientific process hierarchy. It was also established that hypothesizing was a skill associated with the attainment of formal operational level of cognitive functions, a second-order skill in the learning process, and a complex process subsequent to the making of inferences in the hierarchy of scientific processes.

The designed empirical study was to address two issues. One was the validation of the tests developed for the study. The other was to focus on the ability of junior high students to make inferences and formulate hypotheses. The tests were designed from a model proposed by F. Micciche and modified to suit a pencil and paper format. A test of the use of selected scientific processes, the *General Science Test (GST)*, was developed to provide for the

concurrent validity of the *Inference Test (IT)* and *Hypothesis Test (HT)*. The *Cooperative School and College Ability Test (SCAT)* was used to provide information about the student sample and to provide further information about the concurrent validity of the cognitive level of *IT* and *HT*. In June of 1972, 960 students in three jurisdictions in central Alberta were tested to provide the data base for the study. Data analysis procedures provided means, variances, item correlations for factor analyses, step-wise regression analyses, analyses of variance to provide insights into the two main issues and to test the hypotheses about student performance.

Evidence was produced that shows that students improve in their ability to make inferences and formulate hypotheses as their intellectual capacity develops. As is often the case with adolescents there were wide ranges in the item response data that are probably related to the cognitive development in adolescence. Boys performed slightly better than girls on the *IT*, *HT*, and *GST*. Both sexes found the *HT* very difficult which would indicate that large numbers of the students were unable to exhibit a competency related to formal operations.

In general, both from the theoretical rationale and from the empirical study, the concept of the stages of intellectual growth, the cyclical nature of learning and the hierarchical relationships between inferences and hypotheses were supported. There appears to be a substantial discrepancy between the age range for the attainment of formal operations reported in the literature and the performance of students in this study. A number of potentially useful implications for science educators were drawn from the findings of this study.

ACKNOWLEDGEMENTS

In acknowledging all of the encouragement and assistance that I have received over the years since this study was begun would fill a volume many times the size of this thesis. However, I will attempt in a small way to thank those persons whose cooperation has meant the difference between success and failure.

First I wish to express my appreciation — to Dr. M.A. Nay, whose support, encouragement, guidance and understanding helped me to bring the study to its successful conclusion; to the members of my committee for their patience and understanding through the long years; and to Dr. D.W.R. Wilson who, as my initial supervisor, encouraged and guided my first stumbling efforts and whose assistance in the construction of the *Inference Test* and *Hypothesis Test* was invaluable.

I sincerely thank the staff, senior executives and my colleagues in the Department of Education whose continuing support, assistance and patience encouraged me to see the project through to its conclusion.

A very special thanks is gratefully extended to my wife, Carol, and daughters, Maurine, Shauna and Colleen, for their help in marking and recording and for their unstinting support and cooperation during the writing of this report.

M.R.T.

TABLE OF CONTENTS

Chapter | Page

Chapter Page

IV (Cont'd)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION TO THE STUDY

This chapter is an introduction to and a description of the nature of the study. Chapter II consists of a review of the literature for the theoretical framework for the study. Chapter III includes a description of the design of the experiment and of the statistical procedures used in analyzing the data. The results of the analysis are reported in Chapter IV. Chapter V is devoted to a summary of the investigation along with conclusions, limitations of the study, implications for science education and implications for further research.

Background of the Study

The processes of science have become a major theme in the development of science curricula at all levels in the past few years. Many of the courses such as the *Science Curriculum Improvement Study* (*SCIS*), *Elementary Science Study* (*ESS*), *Science — A Process Approach* (*S-APA*), Nuffield Foundation programs, etc., have taken as their rationale the concept that science is more than a collection of facts — it has also a structure and a way of discovering new knowledge. In accordance with this rationale, these courses have tried to give students a feel for science as a whole and to show how scientists work, the kinds of problems they attack and the kinds of thought strategies required to find solutions. One implication for developing a course around these skills and strategies is that any student can learn any subject in an "intellectually honest" form and is able to apply the generalizations and understandings that

have been learned (Bruner, 1961). The structuring of such a course involves the selection of the appropriate content that illustrates the structure and the processes of the particular branch of science being presented.

Most of the National Science Foundation and Nuffield funded science courses that have been developed over the past decade have de-emphasized the passive teacher-pupil lessons in which the teacher told about science in favor of a hands-on, activity oriented program in which the student and teacher do science together. This is most apparent in the shift from text books that act as an information source to those which act as a source of questions to be answered and problems to be researched.

In Alberta, this shift in emphasis has resulted in curriculum guides that emphasize goals that describe an attainment of process skills and an increased reliance on the interaction of the student with concrete materials in the laboratory setting. The Secondary School Science objectives call for a balance of the processes of inquiry and scientific content (Alberta Education, 1974, p.iii). These were modelled to a large extent upon the work of Gagné and others who produced the *S-APA* program.

To summarize briefly, elements from several psychological "schools" are being combined to provide a basis for the focus of this study that is the sequence of observing → inferring → hypothesizing.

Psychological Bases

This growing emphasis on the process skills involved in science thinking raises the question of the ability of students to develop these skills.

The search for a model of intellectual development that includes the development of logical thinking leads one to a consideration of the writings of Jean Piaget. Piaget has devoted a great deal of time and energy to a study of the development of the intellectual skills that we call thinking in a "scientific" fashion.

In describing the development of logical thinking, Inhelder and Piaget (1958) make the point that organization is inherent in intellectual functioning and imposes its structure on thought. They reject the idea that education has imposed this structure on children.

> Society does not act on growing individuals simply by external pressure, and the individual is not, in relation to the social any more than to the physical environment a simple "tabula rasa" on which social constraints imprints ready-made knowledge (p. 338).

They go on to say these structured ways of thinking are the result of the interchange between adolescents and other people, and between adolescents and the physical world. Mature thought patterns are a necessary part of this exchange — necessary but not sufficient in themselves to explain the logical problem-solving behavior of adolescents. In the description of children's cognitive development, Piaget (1973) recounts how the young child at a preoperational stage can deal with concrete objects; the school child in a concrete-operational stage can deal with them in a certain logical fashion — he becomes capable of coordinating operations in the sense of reversibility; the adolescent on entering the formal-operational stage is freed from the bonds of the physical reality to deal with the hypothetical possibilities "of reasoning on propositional, verbal statements" (p. 21). He further makes the point in his emphasis on discrete stages that development is

continuous, and that each stage evolves out of the one before it and contributes to the following one. Children mature at different rates but the sequence of development is the same (pp. 1-30).

There is also a horizontal development of the intellect. With growing maturity, the thought structures tend to be richer, more complex, and more inclusive. Instead of reasoning directly from a particular set of data, many adolescents use indirect, second-order logical operations for structuring the data; instead of merely grouping data into classes or arranging them serially in terms of a given variable, they become capable of formulating and testing hypotheses based on all possible combinations of variables. Since adolescents' logical operations are performed on verbal or symbolic propositions, they can go beyond the concrete and deal with all possible or hypothetical relations between ideas. The formal-operational child can develop second-order constructs derived from relationships between previously established verbal abstractions, already one step removed from the data.

Piaget (1972) holds that an important characteristic of the stage of formal operations is the capacity to relate one proposition to another. This is a characteristic of many adolescents. The adolescent who is thinking in this fashion identifies a problem, determines all of the possible relations; in other words, systematically isolates all of the variables and the possible combinations of these variables. It is by means of this inter-propositional thinking that the organized combinations become testable.

The particular mental structures that an individual can bring to

bear in a problem-solving situation are indicative of the particular stage that he has reached in his cognitive development. These structures become relatively stable over time and, once stabilized, are transferable to most situations. However, before this stabilization occurs, there is some evidence that these *structures d'ensemble* are not generalizable very far from the context in which they are learned. In this regard, Stone and Ausubel (1969) report that "inter-correlations based on scores on tests measuring the application of recently introduced abstract principles from three academic disciplines were consistently higher for tenth graders than for seventh graders of comparable intelligence" (p. 180). Piaget (1972) also comments on this point of the generalizability of formal operations in a discussion of the cognitive evolution of an adolescent:

> At the concrete operations level a structure cannot be generalized to different heterogenous contents but remains attached to a system of objects or to the properties of these objects (thus the concept of weight only becomes logically structured after the development of the concept of matter and the concept of physical volume after weight): a formal structure seems, in contrast, generalizable as it deals with hypotheses (p. 10).

In his model of human intellectual development based on the idea of cumulative learning, Gagné (1968) proposes that new learning depends upon the combining of previously acquired and recalled learnings. He suggests that this transfer and recombination of previous learnings accounts for the increasing sophistication that is observable in individual learners. He also denies the existence of logical structures except in the sense that combinations of prior learnings into new ones carry an inherent logic (p. 189). He continues in the exposition of cumulative learning by suggesting that cognitive structures are generated by the interaction of learned capabilities through learning memory and transfer. This model

fails to explain a child's "intuitive" understandings that follow his experiences with concrete objects. Gagné's description of the sequence of subordinate learnings leading to the successful completion of the "conservation of liquid task" begins with the point, line concepts and progresses through areas of rectangles, taking length and width into account, to volumes, taking length, width and height into account. This model is contrary to that of the Piagetian tradition which emphasizes the need for a great deal of practical hands-on experience with objects such as blocks, containers, sand and water. In his analysis Gagné has depended to a large extent upon a vertical hierarchy of skills, making the assumption that if a child has mastered the prerequisite skills he is ready to learn a skill or concept. In doing so the interrelationships among cognitive structures have been ignored. This "readiness to learn" is more than being ready to learn a linear sequence of facts, skills and concepts but includes the experiential background of the student, the context in which he is to learn the concept. In other words, the sophistication of the concept that can be learned is related to the sophistication of the cognitive structures that the learner has mastered. Gagné's model assumes that learning will take place within a rather narrowly circumscribed domain, and that concepts will not be learned as highly general, but rather as relatively specific entities. He also infers that the subsidiary concepts and skills are the major determiners in the acquisition of scientific knowledge.

Bruner (1973) commenting on the idea of the combination of previous learnings, suggests that creative learning is "not simply a taking of known elements and running them together by algorithm into a welter of

permutations." He continues: "To create consists precisely in not making useless combinations and in making those which are useful and which are only a small minority. Invention is discernment, choice" (p. 210). He then suggests that it is "intuitive familiarity" that leads a learner to the most productive combinations. This "intuitive familiarity" is gained by an individual after close experience with objects and organisms in his environment.

In drawing together the salient features of these psychological "schools" it is not the intention of the investigator to present the various positions in their entirety, but rather to allude to those constructs that seem to be relevant to the present study. For example: *from the Piagetian tradition* — the idea of cognitive growth through a series of stages ending with formal operations, and the capacity to deal with the physical world in an abstract fashion by developing and dealing with propositions; *from Gagné's writings* — the idea of combining previous learnings into new patterns to suit new conditions, as well as his development of the hierarchy of scientific skills and processes; *from Bruner* — the idea of the creative element involved in the combination of previous learning to new, more powerful "inventions" that arise from the individual's "intuitive familiarity" with objects and organisms in his environment.

This is not to say that these ideas have remained identifiably separate or, indeed, that these ideas are uniquely characteristic of each author. There is in fact a fair degree of commonality among these three schools of thought. By direct exposition and by extension of basic premises, it is evident that all three authors believe that

knowledge is gained by experiences with and by operations on concrete objects. From these experiences and operations, individual observations are taken, some conclusions, answers or inferences are made, and ultimately a hypothesis is formulated. There is a degree of agreement that there is a growth in the calibre of logical manipulation of ideas that is evident as thought patterns or mental structures mature. Whether the logic resides in the subject matter being learned or in the learner himself is a point of difference. There is also a degree of commonality with respect to the hierarchical nature of the learnings and skills, but again, whether the focus of the hierarchy is in the learner's mental structures or in the nature of the science he is learning, is a point of argument.

There are also some substantial differences that are readily identifiable: The contention by Piaget that the pattern of intellectual development is fixed and that the rate of development is only minimally modifiable is a point of dispute. In fact, Piaget is not concerned with teaching specifically to affect either the rate or the order of development. Gagné is concerned with having the child learn the skills, processes and concepts of the real world in such a way that the logical structure of the organized body of knowledge becomes the mental framework into which new knowledge can be integrated. Bruner implies that the mental structures are inherent in the learner and that combinatorial skills are taught to enable this "intuitive familiarity" to integrate new experiences and to create new knowledge.

For the purposes of the present study, the combining of observations to form inferences which in turn are accumulated and generalized into

hypotheses is equated with what Piaget has called propositional thinking. In terms of the design of the study, propositional thinking is operationally defined as the combination of the inferring and hypothesizing skills that students exhibit. The hierarchical relationship between the making of an inference and the formulation of a hypothesis has been implied by all three authors in their discussion of the development of abstract statements by individuals and has been accepted as a basic assumption for this study.

Problem

The goal of this study is to make a contribution to the developing theory of scientific learning. Current science curricula call upon the student to learn and to practise scientific processes. Among these is the ability to infer and to formulate hypotheses. The purpose of this study is to investigate the propositional thinking abilities of adolescent students, in terms of their inferring and hypothesizing skills. To determine the relationship between these two skills, a test of inferring skills and a test of hypothesizing skills were designed and administered.

Specifically, the study is designed to gather evidence pertaining to the following central questions:

1. *How is the ability of propositional thinking as defined in terms of the ability to use the scientific processes of inferring and hypothesizing distributed among the student population (by age, by grade or by sex)?*

2. *Is the ability to think propositionally related to a student's scholastic ability and knowledge of scientific processes?*

Answers to these questions were sought from a population of junior high school students in central Alberta, with an age range of 11 to 15 years. Data were obtained by administering four tests during a one week period in June, 1972.

Definitions

In general, the meaning of each of these terms and abbreviations is indicated where it is first used, but for easy reference, the following list of definitions is presented at this point.

Structures. In Piagetian terms the idea of "structures" is in terms of the organizing and integrating thought patterns that people develop as they mature intellectually (Phillips, 1969, p. 109). Schwab (1962) defines structure in terms of those concepts which define the substantive domain of the discipline and determine its mode of inquiry. To differentiate between these two senses "mental" structures will be used in the sense of *structures d'ensemble* and "scientific structures" in referring to the concepts, skills, etc., which define the bounds of science.

Processes. Processes refer to those skills and operations which are associated with problem-solving activities. More specifically, scientific processes are those operations that are used to create, use, and communicate scientific knowledge. Examples of scientific processes are observing, inferring and hypothesizing.

Operations. In Piagetian terms "operations" are internal activities of the mind, as opposed to the sensory-motor or physical activities of the body. Characterized by logical thought processes which are reversible, "concrete operations" are concerned with concrete, existing objects and include ordering, serial arrangements, and classification, as well as mathematical operations. Formal operations are second order operations concerned with logical propositions and hypothetical reasoning, based on theoretical constructs as opposed to concrete objects.

Inferring. In the *S-APA* program an inference is defined as "an explanation of an observation" (AAAS, 1968, p. 111). That is, an inference has a direct reference to an observation. Tannenbaum (1969) described inferring in behavioral terms:

> To demonstrate competence in inferring a student should be able to:
> 1. Draw warranted conclusions from observations.
> 2. Identify the important factors in a given set of circumstances.
> 3. Relate an observation to a given conclusion.
> 4. Differentiate between a statement of fact about an observation and a conclusion arising from the observation.
> 5. Recognize that more than one inference may be drawn from a given set of data (p. 135).

In terms of the Piagetian model of cognitive development, inferring is a characteristic of a person in the Concrete Operational stage. Because of its close relationship with an objective referant, a person should exhibit this skill prior to those skills that are associated with the stage of Formal Operations.

For the purpose of this study inferring is defined as the skill of explaining observations made from a given set of data.

Hypothesizing. The term "hypothesis" has been used in a number of different ways and in a number of different contexts. For Piaget, hypotheses are statements of the combinations of various elements that can be isolated from the raw data arising from a problem. The ability of a child to think in these terms, that is to undertake combinatorial analysis, is closely affiliated with the development of formal operational thought. In this sense, then, hypotheses are testable statements about the relationship that exists between or among variables that can be isolated from a given circumstance.

Tannenbaum (1969) describes hypothesizing in terms of behaviors that students should exhibit in a controlled situation:

> To demonstrate competence in hypothesizing a student should be able to:
> 1. Group a number of conclusions into a general explanation of a phenomenon.
> 2. Distinguish between a proposition that is a general explanation from a statement of fact about an observation.
> 3. Identify the important conclusion(s) that support a hypothesis.
> 4. Test a hypothesis by suggesting or designing an experiment (p. 135).

On a different level, the *S-APA* program has used Gagné's definition (AAAS, 1965) of a hypothesis as "a general statement that includes all objects or events of the same class" (p. 159). He continues in the passage to make a clear distinction between inferences and hypotheses. Inferences clearly apply to single observations or sets of observations of a single event.

For the purpose of this study, and on the basis of the essential similarities among many definitions, a hypothesis is defined as a tentative explanation of an empirical relationship among variables in a given

problem situation.

Propositional Thinking. The ability to form abstract propositions about events and to relate one such proposition to another is a characteristic of a person who is thinking at the stage of Piaget's Formal Operations. These abstract propositions are hypotheses and the mental operations that are used to formulate them are termed propositional thinking. Propositional thinking relates to the mental operations that depend upon the formation of statements and hypotheses no longer related to the objects themselves. The propositional thinking that persons from the ages of 11 to 15 use is closely related to the development of the intellect and the sophistication of a person's cognitive structures. For the purposes of this study propositional thinking is defined as the formation of hypotheses from inferences which have been made as a consequence of observing an event or events.

The *Cooperative School and College Ability Test (SCAT)* is a standardized test of the scholastic ability level of the students in the population.

The *General Science Test (GST)* was developed for the purpose of providing a measure of the knowledge level of students of scientific processes.

The *Inference Test (IT)* was developed to measure the ability of students to make inferences.

The *Hypothesis Test (HT)* was developed to provide a measure of the ability of students to formulate hypotheses.

Hypotheses

The design of the study is based on the following null hypotheses:

Hypothesis 1:

There is no significant difference between the mean score on the IT among boys and girls in age categories from 11 to 15.

Hypothesis 2:

There is no significant correlation between the student scores on the IT and age, sex, SCAT score, HT score or GST score.

Hypothesis 3:

There is no significant difference between the mean score on the HT among boys and girls in age categories from 11 to 15.

Hypothesis 4:

There is no significant correlation between student scores on the HT and age, sex, SCAT score, IT score or GST score.

Hypothesis 5:

There is no significant difference between the mean scores of boys and girls on the combined IT and HT as an indicator of propositional thinking and their age category from 11 to 15.

Hypothesis 6:

There is no significant correlation between student scores on the combined IT and HT as an indicator of propositional thinking and age category, sex, SCAT score and GST score.

Hypothesis 7:

There is no significant difference between the mean scores of boys and girls on the SCAT and the school attended, and age.

Hypothesis 8:

There is no significant difference between the mean scores of boys and girls in age categories of 11 to 15 on the GST.

Limitations and Delimitations

Piaget (1964) indicates that there are four main factors that affect children's progress from one cognitive stage to another:

1. Maturation
2. Experience
3. Social transformation
4. Equilibration (self-regulation).

In the present study the variations in the degree to which students have progressed to the stage of formal operations will not be related to these four factors.

There are three properties of Piaget's cognitive model:

1. Each stage extends and builds upon the one before and in turn becomes the foundation for the next stage.
2. Children pass through these stages in the same order, though at a variable rate.
3. Age is only a rough guide to the stage of development of a particular child.

In the present study these properties have influenced the age group studied and the nature of the *IT* and *HT*. The junior high school-aged children were chosen as the population because there should be a reasonable number of students in Piaget's stage of formal operations. In addition, most students should have little trouble in responding to the *IT* since it refers to a concrete operational skill.

It has been assumed that all students have received some training in observing and inferring since these are part of the present elementary science program which was adopted in Alberta in 1968.

It has also been assumed that students have received only a minimal amount of training in the formulation of hypotheses. The actual amount of scientific process training that the students have had was not determined and is, hence, uncontrolled.

The population is largely rural due mainly to the cooperation of the school jurisdictions and hence cannot be construed as being in any sense a random selection of Alberta junior high school students.

The testing instructions and the length of tests were kept as simple and as short as possible. But inasmuch as there was no control by the investigator over the testing conditions, it can only be assumed that the instructions were followed by the teachers.

Significance of the Study

Information gained from this study will be useful in the design of the scientific process dimension of the junior high school science curriculum. Recent (1974) changes in the junior high science program have called for an increased emphasis on the scientific processes with a concomitant reduction in the knowledge dimension. The study of inferring and hypothesis formulation will bring an increased understanding of the present level of competence of students in these skills on the part of the provincial curriculum committees.

The tests developed for the purposes of the investigation should be of value to the classroom teacher in the design of activities and instruments to measure inferring and hypothesizing skills. This by-product of the investigation is of particular importance in the situation where a teacher is using a laboratory-centred approach to the teaching of science since the statement of operational hypotheses, which are specific statements of the more general propositions defined as hypotheses earlier, direct the design of the experiment.

This study will also be of general use to science educators who have been pressing for an increase in the teaching of the skills of the scientific processes. By providing base-line data on the general level of student competence in inferring and hypothesizing from observations, expectations being held by the scientific community of the school program can be made more realistic.

CHAPTER II

REVIEW OF RELATED LITERATURE

In organizing the search of the literature for related papers and studies, it seemed reasonable to categorize into these areas: those writings dealing with the process dimension of science and hypothesizing in particular; those dealing with cognitive growth; those dealing with the ability of children to learn and use specific science process skills; Piagetian studies in secondary schools; and those dealing with the evaluation of process skills.

The Processes of Science

With the increasing emphasis upon inquiry teaching in science, the process dimension of science has assumed a much greater importance than ever before. It has been suggested by many writers that learning and using the processes of science will enable a student to discover many of the fundamental concepts of science for himself. We are also encouraged to believe that, by learning to use the processes of science, the student will uncover knowledge about his environment and become a lifelong active inquirer.

In an attempt to clarify the role of the science processes in the teaching of junior high science in Alberta, the Secondary School Science Curriculum Committee, through its Junior High School Science Ad Hoc Committee, has published a curriculum guide (Alberta Department of Education, 1969). In this publication the curriculum committee has expressed its concern that, ". . . data and concepts are essential

ingredients in inquiry. However, they are not the totality of science; processes of inquiry must be included and balanced against content in importance," and "Science is at one and the same time a body of knowledge and a process of inquiry" (Alberta Department of Education, 1969, p. iii). From this point certain recommendations about the teaching of science were made and specific texts were recommended. The Minister of Education subsequently authorized the text recommendations, and the objectives of science teaching in Alberta junior high schools have been incorporated in the *Program of Studies for Junior High Schools*. In Alberta the processes of science are clearly part of the formal curriculum.

In discussing the learning of science processes, Bruner (1961) makes the statement:

> There are many ways of coming to the arts of inquiry. One of them is by careful study of its formalization in logic, statics, mathematics, the like. If a person is going to pursue inquiry as a way of life, particularly in the sciences, certainly such a study is essential (Bruner, 1961, p. 30).

He then goes on to say that knowledge about the heuristics of discovery is not sufficient, that these heuristics which teach students to investigate phenomena must be used in exercises involving problem solving and pupil discovery. "The more one has practice the more likely one is to generalize what one has learned into a style of problem solving that serves for any kind of task one may encounter" (Bruner, 1961, p. 31).

David Ausubel (1967 and 1968) has countered this emphasis on the centrality of process as being a waste of time since problem solving skills are learned only within a very narrow context and do not seem to be transferable across disciplinary lines. In Ausubel's opinion (1968)

the most efficacious type of guidance is actually a variant of expository teaching that is very similar to socratic questioning. It demands the learner's active participation and requires him to formulate his own generalizations and integrate his knowledge and response to carefully programmed leading questions. It is obviously much more highly structured than most discovery methods. Ausubel also makes the comment:

> . . . as a matter of fact, pure discovery techniques as employed by scholars and scientists could lead only to utter chaos in the classroom; put a young physics student in a bathtub and he is just as likely to concentrate on soap bubbles and on the refraction of light, as on the displacement principle he is supposed to discover. Elementary school pupils in the inquiry training program are shown a carefully prepared demonstration illustrative of a given principle in physics and are then permitted to ask questions answerable by yes or no. Under these conditions pupils are engaged in true autonomous discovery in the same sense that a detective independently solves crimes after a benevolent providence kindly gathers all of the clues and arranges them in the correct sequence (Ausubel, 1968, p. 492).

He quarrels not with the basic premise of discovery learning so much as with Bruner's interpretation that the organizing and creative effects of learning by discovery are attributable to the act of discovery rather than to the structure and organization which was put there by the programmers of such a curriculum.

Gagné (1963) agrees that the learning of process skills is a necessary and vital objective of science instruction. He maintains, however, that if the practice of process skills is to be carried out successfully there are two major prerequisites:

1) a suitable background of broad generalized knowledge which can be used in solving problems to make the inductive leap that characterizes inquiry; and

2) the possession of incisive knowledge which makes it possible to discriminate between a good idea and a bad one.

Gagné (1963) indicates that as a child progresses from kindergarten through college there should be four stages of instruction which would enable the child to become progressively a competent performer, a student of knowledge, a scientific inquirer and an independent investigator. He then describes the kind of program that would occur at each stage. In concluding his discussion, Gagné indicates:

> . . . it must be clear that practice in inquiry for the student of science is of great value. But to be successful, it must be based on a great variety of prerequisite knowledge and competencies which by themselves are learned by discovery, but inconceivably by what is called inquiry (p. 153).

In another paper (AAAS, 1965) he becomes more specific with respect to process skills. He identifies the basic skills, those which form the basis for further science learning, as: observation, measurement, classification, communication, prediction and inference. He then continues in describing a program to teach these skills. In grades kindergarten to three, the child is to be introduced to a variety of content in acquiring these skills. This content is to be derived from more or less familiar objects and phenomena in the world around him. By the time the child has reached the end of the third grade he will have acquired some important fundamental process skills, a good many basic scientific concepts and some knowledge about the natural world. The development of these process skills will be somewhat fragmented and disorganized. In the grades four to six, then, the process skills must be practised in a manner that will demand their integration to insure that they will be generalized to a systematic approach to science

problems. Gagné continues in his description of the higher order scientific tasks for which students must be prepared. Students need to deal with such integrated activities as hypothesis forming, operational definition, variable control and manipulation, experimenting, model building, and interpretation of data. In describing activities in learning the scientific approach, Gagné (AAAS, 1965) describes formulating hypotheses in this way:

> . . . the objectives of such instruction are to make the student capable of formulating reasonable hypotheses. He should be able to distinguish the hypothesis he makes from observations from which it has been drawn, and also from the observations required to test it (p. 66).

The latter requirement implies that the student is able to make operational definitions of the intervening variables which form a part of this hypothesis. As used in the S-APA program, the term 'hypothesis' is a general statement that includes all objects or events of the same class. Hypotheses may be formulated on the basis of inferences made from observations. For example, the hypothesis that *all substances soluble in water will dissolve faster in hot water than in cold water* could result from the observation of a number of sugar cubes dissolving at different rates in water of different temperatures. One inference that may be made about the sugar cubes as they melt is that *sugar cubes dissolve faster in hot water than in cold water*, and after a number of trials with different substances and different temperatures of water the resulting generalized inference, or hypothesis, is that *the rate of solution of substances soluble in water is temperature dependent.*

In attempting to come to some definition of this aspect of science, the process dimension, many authors and investigators have either

developed lists of skills and asked scientists to pass judgment on them or have analyzed the work of investigators and have derived what seems to be a logical set of operations that categorize the activities of many scientists.

In this first area is the work of Nay (1971). In this listing of the processes in scientific inquiry identified by Nay, there are five general and 17 specific processes. There has been an attempt at ordering the processes beginning with identifying and formulating a problem, seeking background information, predicting, hypothesizing and designing the data collection. The listing goes on and is very specific in terms of the operation being described.

As a contrast, Mechner (1965) lists subdivisions of scientific methods and research skills which seem to be manageable pieces of discovery in science in more general terms. He lists such things as deductive reasoning skills, inferential skills, hypotheses generating skills, etc., without any indication of an attempt at developing a logical teaching order.

To develop a useful description of science processes such as that developed by Nay (1971), a group of science teachers in India developed a statement of the processes of science that would enable them to meet the objective "to teach the processes of science." This was reported by Brown (1968) and in the description he has listed a hierarchy which conforms roughly to the scheme of Bloom's Taxonomy, i.e., in an ascending order from simple to complex. However, these are not in a sequential teaching order but are a logical sequence of operations that have been developed by observation.

It would appear then that science processes have been identified as being operations designed to use existing scientific knowledge in the search for new information about the universe. It is also clear that these processes are an important part of science for a student to learn and use in his search for understanding. It is equally clear that these processes can be operationally defined and ordered so that teachers can come to grips with them.

Hypothesizing as a Scientific Process

In a survey of 22 state science curriculum guides reported in 1961, Milton Pella noted that there was no mention of hypothesizing as a desirable goal. He goes on to suggest that perhaps the most creative aspect of the scientific enterprise is hypothesis development and the development of techniques for testing hypotheses (Pella, 1961). He also makes a comment to the effect that the benefits of the laboratory are expanded to contribute to several of these creative objectives of science by including hypothesizing and hypothesis testing.

Eric Rogers (1960), in a book designed to teach physics by a studying, experimenting, reasoning, self-study approach, gives hypothesizing a very central role in the generation of new knowledge. He also notes that the terms 'theory' and 'hypothesis' have become vague in general use and are almost always confused with each other. He then goes on to define them:

> Hypotheses are single tentative guesses — good hunches — assumed for use in devising theory or in planning experiments, intended to be given a direct experimental test when possible. Theories are schemes of thought with assumptions chosen to fit experimental knowledge, containing the speculative ideas and general treatment that makes them grand conceptual schemes (Rogers, 1960, p. 343).

In a science teaching methods text, Thurber and Collette (1964) make the comment:

> A large part of scientific thought is devoted to formulating explanations and planning methods for testing these explanations. These processes make up the very heart of what many people choose to call the scientific method. Certainly no young person has a knowledge of science if he is not well skilled in dealing with hypotheses (p. 89).

The authors then continue with a description of the derivation of a hypothesis from data and the subsequent testing and modification of the hypothesis.

In the AAAS science program teacher's guide, *Commentary for Teachers* (AAAS, Xerox, 1970), integrated science processes are described beginning on page 122 in terms that a classroom teacher can use in planning lessons. These integrated processes are: 1) formulating hypotheses; 2) defining operationally; 3) controlling variables; 4) interpreting data, and 5) experimenting.

In *Science — A Process Approach* (AAAS, Xerox, 1970), a hypothesis is defined as a generalization that includes all objects or events of the same class. Hypotheses may be formulated on the basis of observations or of inferences. An example of a hypothesis that is generalized from an inference is as follows: if you invert a glass jar over a burning candle, the candle will continue to burn for a short time and then go out. From this experiment, the following observations can be made: candle burning; candle covered; candle continues to burn; candle slowly goes out. These lead to the following inferences: (1) whatever the candle needs to continue burning has been used up; (2) the candle becomes too hot so burning is no longer the reaction; (3) there is a build-up of smoke and other products so that the flame is smothered.

A hypothesis based on this inference might be that that *burning candles covered with glass jars go out eventually either through removal of a fuel, removal of a condition necessary for combustion, or the addition of some inhibiting element.* To paraphrase then, a hypothesis may be thought of as a generalization based on a number of inferences all of which are the direct result of an observation (pp. 149-150).

In describing the role of hypothesizing in science teaching, J. H. Woodburn (1969) makes the comment:

> For various reasons, the hypothesis phase in episodes of science and in science lessons seems to predestine the success or failure of the scientist or science teacher . . . For the science teacher it is highly improbable that his students will move effectively into the data gathering phase of any lessons unless they appreciate fully the hypothesis or hypotheses for or against which the data are to be marshalled (p. 333).

He continues, "hypothesizing is a very satisfying and enjoyable phase of scientific investigation because of the opportunities to be creative and inventive." For Woodburn then, the hypothesizing activity is a very crucial element in the success or failure of a given lesson. He defines hypothesis metaphorically, "hypotheses are the temporary bridges that a scientist constructs along his mental pathway between initial curiosity and later acceptable understanding" (p. 335).

Other writers seem to fall into one of two categories in their view of hypotheses. Postman (1957, pp. 249-258), Bruner (1956, p. 129), Byers (1965, pp. 337-342), and to some extent Piaget, tend to agree with the role of a hypothesis as an organizing and sorting construct. Science educators and scientists have tended to add the element of explanation to that of selection. Alpren (1946), Atkin (1956), Fredriksen (1959), Rogers (1960), Gibbs (1967), Barker (1970), and Quinn (1972) all define

hypotheses as tentative explanations of observations assumed for use in devising an experiment.

For the purpose of this study, a hypothesis is defined as an explanation of an empirical relationship among variables in a given problem situation. This relationship is inferred from a number of observations which will be given in defining the problem situation.

Comparison of Brunerian and Piagetian Theories of Cognitive Growth

Investigations by Bruner and his associates into the thinking processes of adults were described in *A Study of Thinking* (Bruner, Goodnow & Austin, 1956). This team was concerned with the means by which concepts were attained:

> . . . the search for and testing of attributes that can be used to distinguish exemplars from nonexemplars of various categories, the search for good and valid anticipatory cues . . . [as distinguished from] concept formation [which is] the inventive act by which classes are constructed (pp. 232-233).

They dèpended upon the assumption that ". . . virtually all cognitive activity involves and is dependent on the process of categorizing" (p. 246). These investigators devised tasks to facilitate investigation of the *strategies* used by adults in attaining certain concepts. In one task, subjects were given an array of 81 cards exhibiting one, two or three figures, the figures being crosses, circles, or squares, having one, two or three borders and being colored green, black, or red. The subjects were told that a "conjunctive concept" meant a set of cards that share a set of attributes. Some practice examples were given. The subjects were then shown a card that illustrated a given concept.

Then, from the rest of the cards examined one at a time, the subjects were to indicate whether or not the card was an exemplar of the concept. Each subject could make one hypothesis concerning the concept after each choice attempting to arrive at the concept as efficiently as possible. Four different strategies were identified:

1) a *simultaneous-scanning* strategy in which the subject used the results from each choice to deduce if a hypothesis is tenable or not;
2) a *successive-scanning* strategy in which a single hypothesis is tested with each choice and in which the subject limits his choices to instances providing a direct test of his hypothesis;
3) a *conservative-focusing* strategy in which the finding of a positive instance is used as a focus in making a sequence of choices, each of which alters only one attribute of the focus card to see whether a positive or negative instance is generated; and
4) a *focus-gambling* strategy in which a positive instance is used as a focus, but more than one attribute is changed from choice to choice.

Similar experiments were devised to investigate the attainment of "disjunctive concepts," defined by cards exhibiting a specifiable relationship between the defining attributes (Bruner, et al., 1956, pp. 41-43, 83-89).

Bruner found that there are six basic essentials to concept attainment tasks:

1) an array of instances characterized by observable attributes to be tested in order that a concept be attained;

2) with each instance tested a person makes a tentative hypothesis (tentative prediction) about some combination of attributes that forms a concept, i.e., an abstraction of his experience;

3) with each succeeding instance the hypothesis is tested;

4) each hypothesis and test results in a further limitation of the relevant attributes;

5) the sequence of decisions based on hypotheses and tests forms a strategy for discovering valid cues; and

6) individual decisions made by the person during the sequence are seen as being important in the ultimate success or failure in attaining a valid concept (Bruner, et al., 1956, pp. 233-234).

One important finding from this concept attainment study is that it is possible to describe a cognitive strategy by means of a number of discrete steps and that these steps will vary according to the perceived need for efficiency. The conservative focusing strategy is practically failure proof, cognitively unstrenuous but not very efficient. When under some time pressure, subjects tended to favor the riskier but faster strategy of focus gambling. The researchers also found that the tendency to depend on relevant cues from previous situations may be helpful but also may form a major obstacle for the adoption of newer, more efficient strategies. It was found that there was a general tendency for subjects to be unable or unwilling to use information based on negative instances or indirect tests of hypotheses. It was also found that greater amounts of information were required before subjects would abandon hypotheses that fit their preconceived notions (Bruner, et al., 1956, pp. 235-238).

Since 1956, Bruner has become increasingly interested in cognitive development. The influence of Piaget on his work has been acknowledged publicly in some of his more recent writings (Bruner, 1973, pp. xiii-xxiii). The theories and descriptions of Bruner and Piaget do have points of disagreement, but these are relatively minor in comparison to the fundamental agreements.

The Brunerian view of cognitive growth depends largely on the mastery of techniques transmitted by the culture. To make sense and to learn from the recurring patterns in the environment, a person must be able to represent them in some way. Retrieval and use of relevant experiences depends to a large extent on how these experiences are processed and coded. Bruner has postulated three modes of representation by which individuals construct models of the real world. These are: *enactive representation*, where the individual summarizes events by means of appropriate actions; *iconic representation*, where the individual summarizes events by selectively organizing perceptions and images by spatial, temporal and qualitative structures; and where a person exercising *symbolic representation* represents objects and events by means of arbitrary symbols.

These three modes of representation are operative in the growth of human intellect and their iteration is central to growth. At the enactive level, actions cannot be transformed, and children thinking in terms of actions merely perform one action after another, sometimes reordering them. On the other hand, images, the basis of iconic representation, can be transformed but they lack generalizability. Once a child has learned to handle symbols though, it may well be that

he can use actions and images quite arbitrarily, as symbols, and may use all three modes of representation simultaneously. Bruner (1973) goes on in discussing cognitive growth by stating: "Growth involves not a series of stages, but rather a successive mastering of three forms of representation along with their partial translation each into the others" (pp. 316-317).

In discussing teaching methods, Bruner (1973, p. 418) suggests:

1) that learning cycles can be regarded as microscopic copies of the developmental cycle. He is reported as having written:

2) . . . a teaching method that takes into account the natural thought processes will allow the child to discover . . . by giving him an opportunity to progress beyond his own primitive mode of thinking through confrontation by concrete data,

3) and continues to stress the importance of teaching "so that the student grasps the structure of a subject."

Piaget and Inhelder would want to qualify this point of view to insure that the *student* actively develops his own notions of the structure of the subject rather than being "taught" the structure in meaningless isolation from what it organizes.

In regard to one of Bruner's more well-known remarks: ". . . any subject can be taught . . ." Piaget comments that Bruner overlooks the biological character of development (Piaget, 1966, p. 318). But Bruner has used Piaget's own theory in defence of his view by urging that curricula be spirally designed beginning with concrete ideas being gradually re-presented and developed as the child's thinking processes mature (Bruner, 1973, pp. 423-425).

There is, then, substantial agreement on the developmental nature

of the intellect, but there is also some disagreement on the limitations that this places on the ability of the student to learn scientific concepts and skills. Bruner asserts that readiness is only a half-truth because one can "teach" readiness by providing opportunities for its nurture rather than simply waiting for it to develop in due course. Piaget first dismissed this emphasis on speeding up development as "the American question" but would suggest that there are limits to the amount of "preparation" that one can give a child. Indeed, Inhelder is reported as saying with respect to scientific experiences in grades one and two, that

> The effect of such an approach would be, we think, to put more continuity into science and mathematics and also give the child a much better and firmer comprehension of the concepts which, unless he has this early foundation, he will mouth later without being able to use them in any effective way (Bruner, 1973, p. 420).

Piaget's theory, using such constructs as schemata, assimilation, accommodation, operations, reversibility and equilibration, appears to be quite adequate to account for the effects of past learning and purposive behavior. Piaget's insightful descriptions of cognitive development and his identification of the factors influencing such development are replete with implications for science in education. For example: concepts should be built upon suitable concrete experiences; students can be led to create their own notions of the structure of a subject if they are given a rich experiential background; and a more flexible approach to problem solving is encouraged by having a student discover the interrelationships involved by encouraging a use of combinatorial analysis, allowing students to approach problems from several points of view.

Bruner's description of concept development has also been found to be insightful and applicable to the classroom. The notion regarding recurring learning cycles which carry a learner to higher and higher levels of abstraction and generalization have been related to Piaget's notion of cognitive growth.

Adolescent Learning of Scientific Process Skills

Between childhood and adolescent, thinking processes become less bound to concrete experiences and more dependent upon abstract reasoning and manipulation of hypothetical relationships. This has been observed by many researchers following the lead of such people as Piaget, Inhelder, Bruner, Ausubel and Gagné.

Piaget makes a distinction between the problem of development and the problem of learning. The development of knowledge is spontaneous, tied to the development of the nervous system and of mental functions. Development determines a certain totality of structures of knowledge (a totality of possibilities and impossibilities) for each person at each stage of growth. Learning is provoked by situations as opposed to being spontaneous and is limited to a single problem or structure at any moment. In learning, a particular social environment is indispensable for the realization of an individual's mental possibilities, and such realization can be accelerated or retarded by the nature of cultural and educational conditions. In Piagetian terms, each element of learning occurs as a function of total development as opposed to the view that development is the cumulation of a series of specific learned items (Piaget in: Ripple & Rockcastle, 1964, pp. 7-8).

The central idea in the development of knowledge is that of an *operation*. An operation is an internalized action which can modify an object of knowledge. An operation could consist of constructing a classification by joining objects in a class or set, or of putting events or objects in a sequence.

> To know an object is to act on it. To know is to modify, to transform the object, and to understand the process of this transformation, and as a consequence to understand the way the object is constructed (Piaget in: Ripple & Rockcastle, 1964, p. 8).

An operation is reversible in the sense that it can take place in both directions (being joined and separated). The attainment of reversibility requires more than an ability to undo a transformation (*renversabilité* for Piaget) in that the individual must anticipate in thought a return to the prior state. This anticipation characterizes the attainment of reversibility (Inhelder & Piaget, 1958, p. 35). An operation is always part of a total structure and is linked to other operations. For example, a class exists in a total classification structure; an event is in the context of a sequence which constitutes structure (Ripple & Rockcastle, 1964, pp. 8-9).

In a Piagetian approach to cognitive development it is axiomatic that a given stage is properly understood in the context of the earlier stages. An understanding of an adolescent's thought patterns then begins with a description of the preoperational child, that is one who operates largely in terms of the phenomenal, before-the-eye reality. The concrete-operational child is one who begins to extend his thought from the actual towards the potential (Inhelder & Piaget, 1958, p. 248). This development is a natural consequence of the formation of concrete-operational structures. The example that Flavell uses is:

> . . . looking at a concrete series of three seriated elements $A<B<C$ (the actual) (the concrete operational child) is much more disposed than the preoperational child could be to anticipate the extension of the series to new as yet unordered elements D, E, etc., the potential (p. 203).

The concrete-operational child has several limitations in his ability to conceptualize (Inhelder & Piaget, 1958, pp. 219-251; Flavell, 1963, pp. 203-204). The starting point for concrete operations, as for preoperations, is always the real rather than the potential -- any extrapolation to the potential is seen as a special case activity. He cannot delineate all the possible eventualities and then test for the reality of these possibilities. A second limitation arises from the fact that the concrete-operational child is bound to the phenomenalogical here and now. Each variable must be treated independently, his cognitive development has not progressed to the point where he can deal with a context-free, once-for-all, structuring of information. For example, Flavell cites that the understanding of conservation of mass may be achieved quite independently of the conservation of weight and volume even with the same objects (Flavell, 1963, p. 204). A third limitation lies in the independence of learned logical groupings; they do not interlock to form a simple, integrated system. The concrete-operational child possesses two kinds of reversible operations, negation and reciprocity, but does not possess a total system which permits him to coordinate the two and solve multivariate problems.

The most important general property of formal-operational thought is that reality is conceived as a special subset of all the possibilities which arise from a situation. This orientation implies a strategy that can determine the reality in a set of possibilities, that is what

Piaget calls "hypothetic-deductive" in character. The adolescent now can construct new operations — operations of propositional logic as opposed to operations on objects or events. The adolescent performs these first-order operations but he then continues, after forming hypotheses about the relationship between these first-order results, to operate at a second level. Flavell (1963, p. 206) describes an adolescent who "confronts a problem . . . to determine all the possible relations [and] . . . systematically isolate all the individual variables plus all combinations of these variables." He calls this combinatorial analysis. In this paradigm, then, these combinations are regarded as hypotheses some of which will be confirmed by subsequent investigation.

While a concrete-operational child can test the effects of a variable by negation, removing of the variable from the operation and, by reciprocity, holding the variable constant while studying a second factor, the adolescent can negate and/or neutralize one variable, not simply to study that one variable but to study the action of some other variable — "A transition towards genuinely scientific methods of analysis" (Flavell, 1963, p. 210).

Piaget reports that there is research to show that at between 12 and 15 years the individual starts to carry out operations involving combinatorial analysis independent of school training. Specifically, when given five bottles of colorless, odorless liquid, three of which combine to make a colored liquid, the fourth is a reducing agent and the fifth is water, the student can discover the generalization after having worked out all of the possible ways of combining the liquids

(Piaget, 1972), p. 5). It is the availability of this combinatorial system of thinking when applied to propositions that is one of the essential features of formal thought.

The application of combinatorial analysis to propositional logic for two propositions, p & q and their negation, implies not only the 4-base associations (p and q, p and not q, not p and q, not p and not q) but also the 16 combinations that result from linking these base associations 1 to 1, 2 to 2, 3 to 3 plus all 4-base associations and the empty set. This propositional logic depends on the fundamental propositional operations of inclusive disjunction, implication and incompatibility.

An example of the use of these operations is as follows:

> If the identity transformation (I) is p > q, then its negation (N) is N = p and not q. This proposition can then be changed into its reciprocal (R) such that R = q > p, then its correlative statement (C) is C = not p and q. That is, I = (p, q)
>
> N = (p, not q)
>
> R = (np, nq)
>
> C = (np, q).

The commutative 4-group is:

CR = N, CN = R, RN = C and NRC - I.

The last statement combines in one operation the negation and the reciprocal which was not possible at the level of concrete operations. A physical example of this is the relationship between a moving object that can go forward and backwards (I & N) on a board which itself can

go forwards or backwards (R & C) in relation to an external reference point. Piaget thinks of the INRC group as a model of adolescent cognition, a model based on developmental experiments. The formal-operational adolescent identifies all four transformations and sees them as constituting a system which can then be used to explore possible solutions. Inhelder and Piaget (1958) describe the adolescent as living in the present and in the non-present, that is, in the future, in the spatially remote, and in the hypothetical. They conclude that ". . . logic is not isolated from life; it is no more than the expression of operational coordinations essential to action" (Inhelder & Piaget, 1958, pp. 341-342). In other words, the adolescent has become capable of reflective thinking and his thought makes it possible for him to escape the concrete present toward the realm of the abstract and the possible. He becomes able to conceive hypotheses prior to experimentation as propositions for empirical test.

For example, the formal-operational adolescent can pose a statement like: "If a weight is moved further away from the fulcrum of a balance then more weight must be added on the other side to keep it level." This hypothesis can easily be put to the test, but it is not so simple to formulate the hypothesis in the first place. To do so involves a narrowing of the general problem (of equilibrium in the balance in this case) and proposing an answer to a question — *How can the balance be kept level when the weight is moved further from the fulcrum?* — before the investigation begins. This is a skill that a concrete operational child is generally unable to do; he learns mainly by "seeing what happens" rather than by stating possibilities and testing to see which corresponds with reality. He tends to stay with the "here and now" and not speculate about the "possible."

One major problem with this simple clear-cut model is that it is neither simple nor clear-cut; there is no clear division between the point when a child does not make hypotheses and the point where he does. There is a gradation from the situation where the variables are few and easily defined, to the more complex situation of multiple variables. Progress at this age is indicated by a gradual shift of emphasis in the importance of a hypothesis.

According to Piaget, the development from one set of mental structures to another is explained by the influence of four factors: maturation, experience, social transmission and equilibration. He states that none of these is sufficient by itself to account for the developmental sequence but he considers equilibration or self-regulation to be the fundamental factor (Ripple & Rockcastle, 1964, p. 10).

Even though *maturation* of the nervous system plays an indispensable role in development, it is not a total explanation since the average age at which each of the various stages of development occurs varies widely from society to society.

Experience with physical objects is also a basic factor in the development of cognitive structure, but again it provides only part of an explanation. For example, conservation of mass becomes fixed for a child at about age eight, but he does not assert that weight or volume is conserved until some time later. Weight and volume are describable properties of matter, but how can the amount of substance or mass be considered in isolation of weight or volume? How can a child understand that there was a transformation of the shape of a quantity of plasticene? Something must be conserved because the transformation can be reversed and the plasticene can be returned to its original condition. Since neither weight nor volume is seen to be conserved, the idea of

conservation of mass is merely a logical necessity — no experience can show a child at this level that there is the same quantity of material. It becomes clear, then, that there are two kinds of experience involved, *physical* and *logico-deductive*. *Physical* experience conforms to our usual notions of acting on objects and gaining some knowledge about the objects through the process of abstraction. *Logico-deductive* experience is drawn from the *actions* effected on the objects. For example: a child discovers that, no matter how he groups and arranges a given set of pebbles and no matter in what direction they are counted, he always has the same number. To count, order, or classify the pebbles, action is necessary. The child discovers that the action of counting is independent of classifying; this is a property of the actions, not the pebbles. This is the beginning of the logico-deductive mode, which is further developed by the internalization of the actions so that they can be combined without the need for pebbles. Before the formal operations stage, the coordination of such actions requires the support of concrete objects; but it later leads to logical structures in which operations are combined through the use of symbols and earlier mental structures are used as points of departure in developing new combinations. The source of logic lies in the coordination of such actions as ordering and classifying. Logico-deductive experience, an experience of an individual's actions, is a necessary precondition before there can be operations (Ripple & Rockcastle, 1964, pp. 11-12).

Social transmission is a third basic factor: Piaget's observation that the emergence of formal thinking corresponds to the age at which society expects a child to begin assuming an adult role and not the onset of puberty or other physical change in the child. The distinctive feature of adolescence in modern cultures is the pressure on the child

to assume a new role (Inhelder & Piaget, 1958, pp. 335-336). However, for a child to receive information from society he must have a structure that enables him to assimilate the information. Consequently, social transmission by itself is not adequate to explain development.

Equilibration serves to relate the other three factors. An individual engaged in the act of knowing is led to react to compensate for external disturbances so that a state of equilibrium can be reached. The process of equilibration leads to operational reversibility which is characterized by a dynamic equilibrium in which a transformation in one direction is compensated for by a transformation in the opposite direction. This active process of self-regulation embodies the concept of feedback from the individual's interactions with his environment, and it takes the form of a succession of levels of equilibrium. Levels of equilibrium can be identified according to the probability of the occurrence of various possible forms of compensation. Laws of equilibration determine, at each stage of development, the best forms of adaptation commensurate with maturation, experience and social milieu. For example, the preoperational child can only cope with one dimension at a time and is led to assert nonconservation of a substance whose perceived form is altered, whereas a child in the concrete operation stage is able to take account of compensating changes in dimension to arrive at the assertion of compensation. In doing so he is able to focus on the transformation and not on the final configuration.

In rebutting criticisms of the idea of equilibration, Piaget (1961, pp. 279-281) talks of equilibration as a causal idea which permits the explanation of changes in thinking forms by means of probabilistic schema. In addition, the process of equilibration is characterized by sequential control with increasing probabilities. It starts at the

level of self-regulation and sensory-motor feedback and leads to operational reversibility and intelligent thought at higher levels of development.

So far Piaget's view of the development of cognitive processes has been presented, but what can be said of his view of the learning process? He maintains (1970) that the learning of logical structures can be accomplished only if the teacher can build the structure to be learned from simpler, more elementary logical structures. This is derived from his view that logical structures are only indirectly the result of experience with physical objects. They are only grasped through the function of equilibration in coping with the characteristics of a number of various actions. The learning of complex structures seems to obey laws similar to those governing the natural development of simpler structures. That is, learning is subordinated to development. Learning is only effective if it can be generalized to new situations and if the learner's operational level is raised. Naturally developed cognitive structures satisfy these criteria, and "learned" structures should satisfy the same criteria. Learning in Piaget's view is possible only when there is active assimilation on the part of the learner, assimilation in the sense of integration of reality into cognitive structures (Ripple & Rockcastle, 1964, pp. 15-18).

There are a number of implications for the study and teaching of scientific processes to students that arise from this model of cognitive development. One is that a child's cognitive development has a number of identifiable stages each with its related characteristic competencies. They range from the earliest ability to react to specific objects in the environment to the ability to state hypotheses, analyze the variables and identify interrelationships.

A second implication for scientific education is that the scientific

experiences a child is given at any age should be experiences he is ready for in terms of the stage of mental growth he has reached and they should help prepare him to advance to the next stage.

A third implication arising from this presentation is that, before introducing a child to a new concept, one should test him to see if he has the prerequisites for forming the concept and, if not, he should be provided with the appropriate developmental experiences and thus assist him in the process of accommodation which will lead him to cope adequately with the situation at hand.

A fourth implication relates to the goal of flexible thinking. Since flexible thinking is based on reversible operations it would seem beneficial to teach skills in inverse pairs and to stress their relationship. For example, in teaching classification, a class of objects defines a set and gives that set its identifiable characteristics. The inverse operation is used to define the individual memberships in terms of the identified characteristics.

A fifth implication arises from the nature of combinatorial analysis. This involves putting information together in differing combinations which in turn encourages mental growth by providing opportunities to see things from many points of view. Since mental growth is associated with the discovery of invariants, a systematic search for the features of a situation that remain unchanged over a number of transformations should aid in developing an awareness and understanding of the relationships involved in the situation.

A sixth implication relates to the onset of formal operations at about age 11 or 12. It would appear to be logical to introduce some elements of deductive reasoning in grades 6 or 7 when students can begin to build the mental structures necessary for deductive thinking, in

which the student can then pose hypotheses and propose experiments which enable him to collect and organize data in such a way as to solve the general problem presented.

Peel (1965), a British educational psychologist, has identified one of the more fundamental aspects of cognitive growth during adolescence as a change from "describer" to "explainer" thinking. He sees three noticeable aspects of growth from partial and circumstantial observation to explanatory thought: "(1) comprehensive judgments involving imagination and possibilities, (2) successful use of imagined hypotheses, and (3) spontaneous elimination of less applicable alternatives" (p. 178). These observations by Peel with respect to the young learner have been echoed by other studies and other authors.

Ausubel (1964, p. 261), speaking of Piaget's developmental stages in an article prepared for the Piaget Conference in March 1964, agrees substantially that the adolescent has reached a level of intellectual development where he can formulate and test hypotheses based on all possible combinations of variables. He also agrees that the acceleration of the development of a child's intellect can only be achieved within the limits of the prevailing stage of development. He concludes ". . . one can, at best, take advantage of methods that are most appropriate and effective for exploiting the existing degree of readiness" (p. 266).

Gagné (1963, and AAAS, 1965) expresses the view that the learner progresses through four levels of competence, developing into an independent investigator. Rather than a progressive development dependent upon the increasing sophistication of a child's intellectual structures, Gagné believes that this progression is dependent upon the teaching and acquisition of investigative skills arranged in hierarchical

fashion (1963, p. 153). In support of the importance of training in the process skills independent of intellectual development, he notes that "a child is not a self-aware, analytical, critical investigator" (AAAS, 1965, p. 21). In his view, the child is very egocentric and incapable of handling many logical operations which are fundamental in using scientific processes without an organized training program.

Whether there is a "natural" progression of intellectual development from preoperational to formal operational, as suggested by Piaget, or whether there is a hierarchy of skills to be learned and taught that evolves from the needs of the subject area, as suggested by Gagné, there is some agreement that adolescent students should be competent in the formulation and testing of hypotheses.

Piagetian Studies in Secondary School Settings

Those studies by Piaget and those following his lead that seemed most relevant are those designed to illustrate contrasts between formal operational thought and concrete operational thought. The 16 experiments described in *The Growth of Logical Thinking from Childhood to Adolescence* (Inhelder & Piaget, 1958) were expressly designed to highlight this difference.

In the experiment in which the subjects were to shoot a ball so that it would hit a given target after rebounding off a cushioned bank, concrete-operational subjects were found to be limited to asserting observed relations and to using the relations to shoot accurately. On the other hand, adolescents seemed to look for the general law from

the outset, forming general hypotheses about the regularities and putting them to experimental test (Flavell, 1963, pp. 347-348).

In another experiment subjects were required to explain why certain objects of various densities and sizes would float or sink in water. The concrete-operational children could identify the class of small-heavy objects as sinkers but they did not arrive at a concept of density and they did not relate the amount of water displaced by the object to its mass and volume. However, the formal-operational subjects were able to eliminate contradictions by casting their explanations in terms of an integrated system of variables. They were then led to assert that a given object floats only if its mass is less than that of an equal volume of water (Inhelder & Piaget, 1958, pp. 20-45).

The amount of bending of a rod under a given set of conditions provided the setting to illustrate the growing skill in reasoning of a formal-operational adolescent. The materials involved and procedures employed made it possible to isolate five variables as affecting the amount of bending of any particular rod: the type of metal of which the rod was made, the amount of weight supported, the length of the rod, the thickness of the rod, and its cross-sectional shape (round, square or rectangular). Most adolescents succeeded in differentiating the five variables. Using combinatorial analysis they systematically tested most or all the variable-present and variable-absent combinations. Concrete-operational subjects could discover some of the variables and they did make some attempts to test their effects but their reasoning lacked the "all-other-things-being-equal" mode to demonstrate the effect of each variable. The tendency to use systematic proof seems to be the special

domain of the formal-operational thought structure (Inhelder & Piaget, 1958, pp. 46-66; Flavell, 1963, p. 348).

The experiment involving colorless liquids, some combinations of which would produce a color, was referred to previously in this chapter. In this experiment, it was found that concrete-operational subjects tested by n and n x n combinations to find out those combinations which would produce the color did not systematically eliminate the variables. In contrast, the formal-operational subjects generated systematic combinatorial tests to eliminate those combinations that were inadequate (Inhelder & Piaget, 1958, pp. 107-122).

These few examples are cited to illustrate that the significant difference between adolescent and pre-adolescent thinking is the presence of propositional thinking in the more mature student. The adolescent's formal thinking structures enable him to get past the errors inherent in limited concrete tests by employing the 16 binary operations of formal logic (affirmation, negation, conjunction, disjunction, implication, etc.) to generate systematic tests to isolate relevant variables.

Lovell (1961) has repeated ten of the 16 experiments described by Inhelder and Piaget (1958). Each of 200 subjects between the ages of eight and 18 was examined individually on a selection of four of the ten experiments, with everyone doing the "colorless liquids experiment." A clinical approach was used and the performance of each subject was graded according to nine stages: one stage of pre-operational thinking, four stages of concrete thinking and four stages of formal thinking. The existence of three main stages as described by Inhelder and Piaget was confirmed, and support was found for the assertion that preadolescents

rarely reach the stage of formal thinking. Indeed, not all adolescents exhibit formal thinking with the ablest showing the earliest grasp of the formal patterns and the least able still thinking in concrete patterns (Lovell, 1961, pp. 143-149).

Student performance from experiment to experiment showed considerable agreement among the levels of thinking displayed. The effect of schooling, where there was some overlap in content with school curricula, was found to be minimal. It seemed to the investigator that instruction seemed to have been of greatest value when the required thinking skills were readily available. "If the power to think at the requisite level is not present, knowledge gained by instruction is either forgotten, or it may remain rote knowledge and be regurgitated when required" (Lovell, 1961, p. 151).

The experiments were found to separate students into fast or slow learners in a variety of school subjects, leading the investigator to agree with Lovell's premise (1961, pp. 149-153) that the types of thinking processes involved in the Piagetian experiments are broadly applicable rather than being only relevant to scientific problems.

Elkind's (1961) replication of Piaget's study of conservation, done with 12- to 18-year-old subjects, centred on the influence of age, sex and IQ on the abstract conceptions of quantity in adolescents. Four hundred and sixty-nine Massachusetts junior and senior high students with a mean IQ of 100.4 were given group tests of conservation of mass, weight and volume in that order. On the basis of the results it was found that 87% demonstrated conservation of mass and weight, but only 47% had abstract conceptions of volume. In fact, only 75% of those in

the oldest age group (mean age of 17.7 years) demonstrated conservation of volume.

According to Piaget's work, the majority of children of ages 11 or 12 should be *ready* to attain conservation of volume because this conceptualization only requires concrete operations and because they have had concrete experiences to form abstract conceptions of mass and weight. However, the age at which the child is ready to grasp conservation of volume is also the age at which formal operations are developing. Elkind (1961, pp. 556-557) proposes that this situation produces new interests which tend to reduce the concern with inductive conceptualization from the physical environment in favor of more theoretical interests. He proposes that the possibility of the spontaneous discovery of the conservation of volume is substantially reduced. The adoption of adult roles also beginning about age 11 or 12 leads the adolescent to be more selective in his choice of experience. This, in turn, would cause many adolescents, though ready, not to attain conservation of volume because their role-choices do not provide the necessary experience.

Similarly, the increased proportion of students attaining conservation of volume with increased age can be partly accounted for by realizing that those students who stay in school have adopted roles that would more likely provide experiences which would lead students to the abstract conceptualization of volume. This role-related effect can also be used to explain the slightly better performance of boys in attaining conservation of volume since boys have traditionally chosen a role in the scientific-technical area (Elkind, 1961, p. 558).

Evaluation of Scientific Process Skills

There has been a call for more precise and more complete evaluation of the student's achievement of the objectives of science-teaching. As initially outlined in a paper by Hurd and Johnson (NSSE, 1960, p. 335), a good evaluation "should give evidence of the student's understanding of the science concepts, . . . and his ability to use his knowledge . . ." The call for an improved evaluation was taken up in an article by Reiner (1966) in which he outlined the challenge to educational evaluators to develop new tests and new approaches to student assessment.

In one attempt at clarifying the issue of evaluating in the process dimension, Munson (1967), in a paper addressed to elementary educators, says:

> The expression "teaching science processes" implies performance criteria, therefore measurement must be in terms of specific tests that are developed concurrently with lesson planning. The objective test may adequately measure knowledge of product but it is no longer the only technique of evaluation. We have to use subjective judgments to evaluate science skills (p. 126).

Munson identifies these process skills as designing experiments, stating hypotheses, recording and using data, drawing inferences, grouping and classifying, predicting and drawing conclusions. He closes with the following:

> It is paradoxical to suggest that it is necessary to be unscientific in order to evaluate pupils' progress in science. I hope, however, that teachers will not content themselves with evaluating only product; process or science skills must be evaluated even if only subjective means are used to do so (p. 130).

One could quarrel with Munson's implication that to be subjective one is being unscientific. Since science is a human endeavor, many

observations are essentially subjective to some extent or other, and subject to experimental error. The response of scientists to this source of error is to minimize its effect as much as possible by standardizing practices. In an analogous situation, teachers can standardize their procedures to minimize their own observational errors and the inter-observer error.

In an attempt to develop a test of a student's grasp of scientific processes, Welch and Pella (1968) reported on the development of a test of the knowledge of scientific processes, the *Science Process Inventory* (*SPI*). This report outlines the procedure followed in developing the test. Briefly, the procedure used was to abstract a list of science processes from six basic references. To be included in the list the element must have appeared in three or more of the six references used. This list was then presented to a panel of 14 scientists and revised on the basis of their suggestions. Items asking about the assumptions, activities, products and ethics of science were developed and the 150 items were then administered to 1283 students. Each item has an agree-disagree choice and the answers are keyed to an indication of the student's knowledge of the process. Total scores are obtained by summing the number of agreements with a standard key.

The *SPI* was administered to high school students in Wisconsin and to a sample of science teachers and scientists. The test appears to be quite usable with students who can verbalize their scientific understandings quite fluently. But the test may not be as valid when used with a relatively unsophisticated group of younger students. The reliability estimate of *SPI* is reported as 0.79 based on the split-half

correlation. The validity was estimated by a comparison with a group of science teachers and a group of scientists — the differences between the groups was significant.

On subsequent editing the *SPI* evolved into the *Wisconsin Inventory of Science Processes* (*WISP*) and is available as a standardized test.

Micciche (1969), Beard (1971), Tannenbaum (1971) and Quinn (1972) developed multimedia tests using film loops, sound films, slides, or special devices, and pencil and paper tests to assess a student's knowledge of scientific processes. Beard's test was aimed at primary school children and was designed to assess a child's ability to measure and classify.

A 35 mm slide sequence illustrating laboratory situations involving basic scientific processes was shown along with a synchronized tape recording which provided the instructions to the student and illustrated the problem to be considered. The student indicated his answer to each question by marking his answer sheet as directed by the tape recording. The two samples of the test having content validity, reliability and discriminating power suggest that such multimedia tests where students can see and hear a laboratory situation could be developed as useful evaluation instruments.

Beard's study is important in that it outlined procedures used to develop a multimedia test for administration to young pupils not skilled in reading and writing. Samples of the test composed of validated items were given twice to 854 pupils in grades 1, 2, and 3. Only two of the six samples tested had a product moment correlation of 0.70 or higher. Thus there is some indication that a test that does not rely on a pencil

and paper presentation could be developed as a useful evaluation instrument.

Tannenbaum's test was developed to assess achievement and diagnose weaknesses in the use of scientific processes in junior high. The test is said to assess a student's knowledge of: observing, classifying, quantifying, measuring, experimenting, inferring and predicting. It is a 96-item, 5-choice test requiring 73 minutes for administration and uses 12 35-mm color slides to illustrate the first 12 questions.

One of the eight processes that form the basis of the test is inferring; a behavior that was tested as an example of inferring was to "identify and specify observations which would be needed to justify a particular generalization." An item from the test which was said to sample that behavior was item 94 from Form II:

> In order to prove that "NOT ALL THINGS GET BIGGER AS YOU HEAT THEM," which of the following would you need to do?
> (1) Find one thing that does not get bigger when it is heated.
> (2) Find all the things that do not get bigger when they are heated.
> (3) Find one thing that gets bigger when it is heated.
> (4) Find all the things that get bigger when they are heated.
> (5) Find all the things that do not change size when they are heated.

The student is called upon to differentiate between the statement that relates to a single observation and one relating to a more general situation. Tannenbaum's test is a useful addition to the tests of scientific processes but it is still heavily dependent upon a pencil and paper presentation — it is still a highly verbal test and as such would tend to measure a student's verbalizing ability as opposed to his non-verbal ability to understand and demonstrate scientific processes.

The test was normed with 3673 junior high students and has a $K\text{-}R_{20}$ reliability of 0.91, 0.91, 0.90 with grades 7, 8 and 9 classes, respectively. The test was validated with a teacher's rating of 35 students being correlated with the students' responses on the test. These varied from 0.115 to 0.477 which would indicate that there is a degree of agreement with the teacher rating of the students' ability to use scientific processes.

Butts (1964) developed tests that are constructed to reveal how a student arrives at a solution. They are based on the assumption that:

> . . . these behaviors are of greatest significance to the practice of science:
> 1. Early formation of a clear hypothesis;
> 2. Specific experimentation with the relevant variables to contrast with random guessing;
> 3. Introduction of control to test the validity of the hypothesis selected;
> 4. Specific attempts at the verification of the hypothesis (1964, p. 118).

To assess these behaviors the tests consist of three parts:

Part I — a description of a problem;

Part II — a series of data or questions which one may wish to use in solving the problem;

Part III — a list of possible solutions, one of which is correct. There are four types of data supplied in Part II, relevant information which is pertinent to the solution of the problem; additional information related but not necessary in the solution; duplicate information which repeats already known facts; irrelevant information that does not lead to the solution. The information is given in a "tab" format so that if the examinee wishes, on the basis of a clue, he may remove the irreplaceable tab. The examiner can then tell which information led to the conclusion.

The five solutions are all quite plausible and attractive to

students. After pulling the tab off an incorrect solution, the student is encouraged to return to Part II to uncover more facts and reformulate a solution.

Scoring is based on the order in which the tabs were pulled. The order implies whether a student has formed a clear hypothesis and is seeking information to test this hypothesis or manipulate variables. Judgments are then made on the quality of the inferred action and are scored on a scale of one to five.

The author has made a further assumption that has not been expressed, that is, that there is a preferred method of solving a problem. A student may exhibit insights into the problem either from previous learnings or intuitive understandings and hence "short-circuit" the testing sequence. Depending on when this understanding occurs the test may or may not record the "process" that the student used to solve the problem. Another problem in administering the test to large numbers of students is in the subjectivity of the scoring and the time involved in the scoring; for the purposes of large-scale testing this test is only of limited value.

Quinn's test consists of 12 of the *Inquiry Development Program* film loops. Each film was shown and then followed by a discussion period in which the teacher's only response was "yes" or "no"; the film was then reshown and a few more questions permitted. The children were then asked to write as many hypotheses as they could in 12 minutes. These papers were then collected and scored using a *Hypothesis Quality Scale*. The test was used with four classes of grade 6 students in and around Philadelphia.

Quinn's (1972) study dealt mainly with the teaching of a skill in generating explanations of given situations, following a very specific format and involving a very clearly identified problem. In essence,

the training was very specific to a given situation and the testing was done shortly after the training. In addition the sample was extremely restricted to four classes in two schools. It would be very difficult to generalize from this study. However, the definitions used and the *Hypothesis Quality Scale* are important contributions to the investigation of this ability.

Micciche and Keany (1969) described an unusual approach to the study of a specific process of science — hypothesizing. The article is a description of a "hypothesis machine" which is a "concrete analogy for indirect observation. . . . Using [Micciche's] apparatus, students can formulate hypotheses to the limit of their imagination" (p. 53). They report that the enthusiasm on the part of students at all levels was very high. The device consists of 15 numbered channels enclosed in a square of transparent plastic raised slightly at one end. The channel walls are not continuous but are interrupted by a clear target area unless they are deflected or captured by the target. On the basis of the observed behavior of the balls, students are to "hypothesize" about the nature of the target. The device has an additional feature, which has both good and bad effects, that is its freedom from specific knowledge concepts, and thus the skills that it measures can be isolated from scientific achievement in the cognitive area. This novelty, though, may be subject to practice effects and perhaps the isolation of a specific element of scientific process may introduce other problems.

These studies have attempted to develop a means of assessing a student's knowledge of the process dimension of science by a variety of means and approaches.

There are other studies and tests such as Atkin (1956), Frederiksen (1959), Gibbs (1967), Mokosch (1969) and Blackford (1970) but these that have been reviewed seem to be illustrative of the pencil and paper and multimedia approaches that have been tried.

The decision was made, on the basis of the review of the tests available, that there was a need for the development of a test of hypothesizing ability. It was further determined that the test should have an element of performance since hypothesizing as a scientific process has such a component. Further, because of the possibility that the students in junior high are divided between the stages of concrete and formal operations, it was determined that the test should be to a large extent symbolic as opposed to being verbal. The other constraint placed on the test was that it be a group test capable of being administered with relatively little training. This last constraint was considered necessary because of the need to obtain valid statistics from test groups which were widely distributed.

Test Development Procedures

Studies and articles that deal with science test construction tend to follow a similar format:

1. Identification of the behaviors to be tested, usually by reference to the literature.
2. Design of the test by "blueprinting" — charting the topics and the cognitive level.
3. Construction of the items.
4. Validation of the test items by reference to a jury.

5. Pilot testing of the test.

6. Revision of the test on the basis of early returns.

7. Administration of the test.

This general outline has been followed by a number of recent studies involving the development of tests such as those by Welch (1968), Beard (1971), Tannenbaum (1971). The jury used for validation varies from three used by Beard and the one used by Tannenbaum to the 19 used by Welch.

A similar pattern of test development is used by the Examinations Development Branch of the Alberta Department of Education outlined in their procedural handbook. The same general pattern is suggested in Wood (1960), Hedges (1966) and Ayers (1967).

This format is considered to be generally acceptable by many writers in the field and, other than large variations in the validation design, studies have tended to follow a similar pattern. The differences in validation appear to be related to the degree of difference from the usual pencil and paper format of multiple choice tests.

Test Validity

The literature on testing abounds with excellent discussions of the characteristic of measurement which is labelled "validity." This characteristic is usually described as "content," "construct," "predictive" or "concurrent."

Content validity usually refers to the correspondence between the test items and the attribute or knowledge being tested. An appropriate technique for checking this correspondence of items with the attribute

being measured involves the use of competent judges. There is bound to be some disagreement over the items; however, Bloom, Hastings and Madaus (1971, p. 76) suggest that 75% agreement or better is satisfactory while less than 50% agreement should be cause for alarm. For this study, six out of nine, or 67%, was deemed to be a minimum acceptable level of agreement.

Construct validity is a characteristic of most ability or personality tests and refers to the relationship of the items that measure the same trait or group of behaviors. A measure of construct validity is the correlation of one item or group of items with others and the extent to which there are common factors within a test.

Predictive validity is a characteristic of ability measures and is a measure of the degree to which the items in whole or in part correlate with either subsequent actions or test performance. This presupposes some sort of logical relationship between the test and a criterion.

Concurrent validity is the extent to which student performance on one test is the same as their performance on some previously established standard. For example, one might suppose that the rank order of students writing a highly verbal test might remain in the same relative position on a verbal-ability test. This is of most use in establishing the relation between an indirect and a more direct measure of some behavior.

CHAPTER III

DESIGN OF THE STUDY

This chapter contains a description of the study population, testing program, tests, validation design and statistical treatments used in the analysis of the total test battery.

Population

The arrangements for carrying out the study were made through contact with the school superintendents of central Alberta. The superintendents from three jurisdictions expressed a particular interest in the study. These jurisdictions, the Counties of Red Deer, Lacombe and Stettler, enrol about 2000 junior high students. From the list of junior high science teachers in these jurisdictions, names were selected by lot until a sample of teachers of about 1250 students was obtained. These teachers were then contacted and all agreed to participate. One teacher asked to be excused after the testing program was begun because of a previous commitment of the class during the test period. The total number of students who participated in the testing program was 1250 in 48 different classrooms and 12 different schools ranging from large 300-pupil schools to a small 15-pupil, 2-room school. The students participating in the study were typically rural students attending schools representative of rural and small town schools across Alberta. Geographically the schools form a broad band across central Alberta from Spruce View in the southwest to Gadsby in the northeast. The majority of the schools are rural centralized schools with large numbers of students bused from their homes.

One school with nine classrooms and 290 students was used to pilot various tests in the battery and the results from this school were treated separately. The total possible number of students for the main study was 960 in 39 classrooms. The distribution of the student population is shown in Table 1. Because of scheduling difficulties, personal and group commitments on the part of students in the study, the number of students responding to an individual test varied from a low of 801 answering the *GST* to 917 responding to the *SCAT*. In analyzing results from the test battery care had to be taken to use procedures that were not affected by the unequal numbers of students. For the final analysis of the test battery, data were included only from the 539 students who had written all four tests. This variation in the size of *N* reduces the generalizability of the study results.

Testing Program

The teachers were asked to administer the four tests in the following sequence: *Cooperative School and College Ability Test (SCAT)*, *General Science Test (GST)*, *Inferring Test (IT)* and *Hypothesizing Test (HT)* (see Appendices A, B, and C) within the same week. To give some flexibility, the tests were delivered during the last week in May and collected during the second week in June.

The *SCAT* form 3B was administered first. This was used to establish the characteristics of the participating classes, that is, to obtain a "reading" on the scholastic ability of the participants of the study. This particular form of the *SCAT* was chosen since it had not been administered with the March battery of the Department of

TABLE I

STUDENTS RESPONDING TO THE INDIVIDUAL TESTS

Tests	Student Sample*		
	Boys	Girls	Total
Inference Test	408	466	874
Hypothesis Test	414	473	887
General Science Test	374	427	801
School and College Abilities Test	428	489	917

*From 11 schools:	Grade	Classrooms	Students
	7	14	298
	8	14	359
	9	11	303
		39	960

Education examinations. Another advantage was the familiarity of the teachers with its administration since it has had a long history of use in Alberta.

Following the *SCAT* form 3B, the teachers were asked to administer the *GST*. This test was designed to measure students' comprehension of a few selected science concepts and their ability to use the higher mental processes. The Examinations Branch of the Department of Education classified the questions in the item bank on a 3-point scale adapted from Bloom's Taxonomy, and most questions on the test were categorized as being a level or two higher. The content-related questions were relatively few in number so that the differing content of the science courses should not bias the student responses to the items.

Following in the testing sequence was the *IT*. This was coupled with an explanatory exercise to provide practice with the format of the test items. This test was to measure a student's skill in making inferences on the basis of indirect observation. The test itself is timed at about 20 minutes with a further 10 minutes for the written introduction and practice items. It fits into a single 35-minute period.

The final test in the sequence was the *HT*. It was also designed to fit into a 35-minute period, including a 10 minute written introduction to the format and the novel format for indicating the answers. The test was designed to test a student's skill at making inferences about a "hidden" object and collecting these inferences into a hypothesis about the object. The respondent was then given an

opportunity to test this hypothesis by extending the given pattern of distribution of balls.

Description of the Tests

In the development of a test of the hypothesizing ability of junior high students several alternatives were explored. One very attractive alternative was to follow the trend to develop Piagetian tasks parallel to the balance problem or the hydraulic press problems presented by Inhelder and Piaget (1958). This was discarded as being too costly in administration and scoring time to be used with large numbers of students and also having the additional inservice problem in training examiners for their general classroom use. The multiple-choice test developed by Tannenbaum (1969) was examined but discarded because of its tie to specific content which may not have been learned by the subjects. The TAB test developed by Butts (1963a) was also examined but discarded because it did not seem to lend itself to machine scoring. In an attempt to reduce the variability due to differences in verbal skills and in learning background, and to increase the level of usefulness, it was decided to try to develop a diagrammatic test that was relatively content-free with a minimum of reading and writing for the student and easy to administer by the teachers.

Inference Test (IT)

In the search for a suitable alternative several diagrammatic tests were explored but it was determined that there were few materials available for use with adolescent students. The EME Hypothesis Machine

developed by Micciche and Keany (1969) was used as a basis for the development of two pencil and paper tests for this study.

The *IT*, included as Appendix C, adapted and modified the Hypothesis Machine model and extended the number and combinations of targets. The targets are blocks and cups of various sizes and slopes of various lengths. This test is to obtain a measure of a student's ability to infer the shape of a hidden target by observing the results of dropping 15 balls down the channels of a device similar to the E.M.E. Hypothesis Machine. The mental process involved was termed an inference because the conclusion or proposition developed is based on concrete data. In addition there is no generalization or testing of the proposition reached, so the conclusion does not meet the definition of a hypothesis used in this study. This is in agreement with Gagné's definition used in the S-APA program, that is, that an explanation of an event or a piece of information resulting from an event is an inference. In Piagetian terms, the intellectual stage being tested is in the realm of concrete operations since the operation being performed is on concrete data and the examinee is not asked to perform operations on the propositions developed.

In the course of the test there is an increase in the complexity of the problems posed but there is no change in the type of operation being asked of the respondents. The test consists of 36 problems to be completed in 20 minutes or less.

Hypothesis Test (HT)

This test is an extension of the basic elements included in the *IT*. Instead of basic geometric forms this test uses more complex shapes

and instead of asking students to merely identify the shape that caused an observed ball pattern they are led through a sequence of changes in the pattern from which they are asked to identify which of the 23 targets is responsible for the pattern.

In the identification of a person's ability to perform operations on operations it was suggested from the literature that, in a given situation, a formal operational adolescent could be expected to explore the possible alternatives, put these alternatives in the form of hypotheses and proceed to test these hypotheses.

The procedure developed was to build upon the *IT* by using a different set of targets and a modified device. The targets are stylized capital letters:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z.

Note that the B, D, O, and Q have indistinguishable patterns, therefore only the O was used from this sequence. Twenty problems were posed using 20 different letters from the 23 available. The letters were used in a scrambled order to reduce the guessing. This precluded an ordering of the patterns from simple to complex. Students were told that the targets were stylized letters and that they appeared only once but they were not told that three were dropped from consideration. To reduce the complexity of the task, the complete list of 26 targets available to the respondents was printed on the test and the respondents were asked to indicate their choice of target for each part of the problem. The letter-target was rotated 90° clockwise to form the next part of the problem until each target was shown in four positions. On the basis of conclusions drawn by inference from the first four parts

of the problem the respondents were asked to hypothesize about the letter-target and to test this hypothesis by indicating the ball pattern that would result from their choice of target.

In Piagetian terms the student is asked to develop a proposition and then perform four coordinated operations in the formation of the proposition: a direct operation (I) and its opposite (N), and in the testing of the proposition the reciprocal of the first (R) and its correlative (NR=C). The main idea being examined in this test is the ability of a student to develop propositions and manipulate them in the solution of a problem. In each of the 20 items the target is presented in its usual orientation, two balls are dropped and the pattern noted. The target is then turned 90° and must be visualized in this orientation. After further rotations of 90° the target is returned to its original orientation. Each pattern must be considered both in isolation and in relation to the previous pattern and the subsequent pattern. 180° rotation has resulted in a reciprocal orientation and a further 180° rotation negates the first rotation.

In terms of paralleling Tannenbaum's (1971) behavioral description the students were asked to:

a) group a number of conclusions (inferences) into a general explanation of a phenomenon.

b) distinguish between a proposition that is a general explanation and a statement of fact about an observation.

c) identify the important conclusions that support a hypothesis.

d) test a hypothesis by suggesting or designing an experiment.

General Science Test (GST)

As part of the design to establish the concurrent validity of the *IT* and *HT* in measuring a scientific process it was determined that a general science test should be administered at the same time. Access to the Alberta Department of Education Examinations Branch's item bank was obtained and a 60-item test was developed using questions that met several criteria. In order to be included, each question was to have: a difficulty of between 0.20 and 0.80, a biserial correlation of more than 0.20 and an item reliability index greater than 0.10. The item bank was searched for items that dealt with the subject area of junior high science, namely, the life, earth and physical sciences, and requiring the use of higher mental processes (Bloom's Application, Analysis, Synthesis and Evaluation categories).

The decision to develop the *GST* using questions measuring level 2 and higher was made on the basis of two arguments: 1) the skills in hypothesizing seem related to a higher level of mental skills as defined by Gagné and Piaget, and 2), the items from the bank that called for some thought and use of tentative explanations or testing of explanations had been classified previously as being of level 2 or higher.

The questions chosen also required the use of selected process skills, that is, they required students to make inferences and hypotheses. The range of processes being tested was deliberately restricted to those related to the formulation and testing of hypotheses.

The initial classification of the items was done by the Examinations Branch committees before they were placed in the item bank. The classification is done in terms of a condensed version of Avital's and

Bloom's taxonomies prepared by the Examinations Branch, consisting of:

1. *Knowledge*:

   To answer items at this level the student needs only to *recognize* or *remember* material learned directly from text books or through classroom instruction.

2. *Comprehension*:

   At this level the student is made aware of the concept that is being tested, but must relate ideas in a *meaningful* way to formulate or identify *new examples or ways of presenting* the concepts.

3. *Higher mental processes and skills*:

   (Application, Analysis, Synthesis and Evaluation)

   From his knowledge and understanding of the subject area, the student at this level must *independently select and use* appropriate concepts and skills which will enable him to deal with a *new* or *unfamiliar* situation presented in the test item. The test item must not state which concept or skill is being tested.

The content areas are life, earth-space, physical and general science. The judgment for inclusion in one of the four categories was made by the item author and confirmed by subsequent users of the items. The blueprint for the *GST* is presented in Table 2. The 60-item test was piloted with 290 students from a County of Red Deer junior high school whose students are bused from the area around Red Deer. The test statistics on this administration were:

TABLE 2

GENERAL SCIENCE TEST: BLUEPRINT OF PILOT VERSION

Mental Activity (Cognitive Process or Level)	Topic or Content Area				No. of Items and Emphasis
	Life	Earth-Space	Physical	General	
Knowledge:				1, 2	2 items 3.2%
Comprehension:	17,35,	7,41 44,54	4, 5, 6, 8, 12,13, 14,15, 16,38, 39,43, 45	3,34, 36,37	23 items 38.3%
Higher Mental Processes and Skills: (Application, Analysis, Synthesis and Evaluation)	21,28, 29,30, 32,33, 47,48, 53,58	19,24, 26,27, 42,46, 49,50, 55,57	18,20, 22,23, 25,40, 51,52	9,11, 31,56, 59,60, 61	35 items 57.4%
Total no. of Items	12	14	21	13	60
Emphasis	19.7%	23.3%	34.4%	21.3%	100%

TABLE 3

GENERAL SCIENCE TEST: BLUEPRINT OF FINAL VERSION

Mental Activity (Cognitive Process or Level)	Topic or Content Area				No. of Items and Emphasis
	Life	Earth-Space	Physical	General	
Knowledge:				1	1 item 2%
Comprehension:	15	6, 8, 36,44	3, 4, 5,10, 11,12, 13,14, 32,33, 35,37	2,28, 29,30, 31	22 items 44%
Higher Mental Processes and Skills: (Application, Analysis, Synthesis and Evaluation)	18,22, 23,24, 26,27, 39,47	19,20, 21,38, 40,41, 45	16,17, 34,42, 43	7, 9, 25,46, 48,49, 50	27 items 54%
Total no. of Items	9	11	17	13	50
Emphasis	18%	22%	34%	26%	100%

Test mean - 24.4

Test variance - 76.0

$K\text{-}R_{20}$ reliability - 0.83

After the piloting and on the basis of the item analysis, it was decided to discard those items which did not meet at least two of the three initial criteria, that is a difficulty level of between .20 and .80, a biserial correlation of greater than .20 and an item reliability index greater than .10. On this basis 10 items were cut and the 50 remaining items were used as a test of higher mental processes in science (see Appendix A). The revised blueprint is presented as Table 3.

The item numbers are different in Tables 2 and 3 because of the deletion of 10 items. The numbers in the blueprint (Table 3) are those of the final form of the test which is included as Appendix A.

The revised *GST* was then administered to 801 students from which the following test statistics were obtained:

Test mean - 21.0

Test variance - 55.8

$K\text{-}R_{20}$ reliability - 0.80

Cooperative School and College Ability Test (SCAT):

(*SCAT*, Form 3B - Cooperative Test Division Educational Testing Service 1956)

The *SCAT* form 3B was administered to the students at the same time as the other three tests. This test was chosen for a number of reasons:

1. It was familiar to the teachers who have administered it as part of the provincial testing program for a number of years, and it is a reliable predictor of success in school.
2. There are provincial norms for the test which can be used to establish the relationship of the sampled students to the general population.
3. The test gives two scores which are used to check the non-verbal nature of the *IT* and *HT*.

SCAT yields three scores: verbal, quantitative, and total. The form 3B is developed specifically for the junior high grades and is administered in a 75-minute time period. The reported reliabilities are considered satisfactory but may be a little inflated because of a speed factor. The test is easy to administer and easy to score partially using the optical score sheets and the scoring program of the Department of Education. At the time of the study, the grade 9 students had written form 3A of *SCAT* three months previously.

Writing in Buros (1959), Fowler reports that "undoubtedly *SCAT* is a superior test which clearly shows the result of careful planning, an excellent experimental program and the use of sound up-to-date statistical procedures" (entry 322, pp. 453-455). He cautions, though, that the test should be only used as a predictor of success in school, which it does exceptionally well, and not as a diagnostic tool. He points out that the validity coefficients are as high and often considerably higher than similar coefficients for other tests of this type. The $K\text{-}R_{20}$ coefficients are at least .95 for the total score

at all levels. For the verbal scores the internal consistency coefficients are at least .92 and for the numerical scores they are .90 or greater. In a subsequent yearbook (Buros, 1965), Green comments that *SCAT* can be "regarded as a set of very good scholastic aptitude tests which probably is in most ways the equal of any of its competitors. In most ways, also, it is a good model of how such a series should be planned, developed, standardized and validated" (entry 452, pp. 717-718). He also points out that the *SCAT* is not useful as a diagnostic tool but is a good predictor of future performance for use from grades 5 to post-secondary level. He concludes with the comment, "It is a good general IQ test from which one cannot legitimately calculate IQ's."

The *SCAT* produced three scores: a verbal, a numerical, and a total score of "scholastic ability." One problem with the test that has led to some difficulty in using the American norms is that the numerical problems tend to be phrased in terms which are not used in Alberta mathematics programs. However, the Alberta norms developed over the past 15 years do provide a basis for comparisons.

Validation of Tests

The validity of the tests developed and used in this study is a crucial question. Considerations related to the validity of each test will be discussed in turn.

Cooperative School and College Ability Test (SCAT)

There is ample evidence of the construct and predictive validity and reliability of the *SCAT* (Buros, 1959, entry 322; Buros, 1965, entry 452) when used to measure scholastic ability and to predict performance

on school-related tasks. One point that Bloom, Hastings and Madaus (1971, p. 78) make in discussing validity is not so much the absolute validity of a test but the validity of the use of the test and the results of the test. In this study the use of *SCAT* is quite appropriate since it is theorized that the inferring and hypothesizing skills are closely related to the reasoning skills that are measured by *SCAT*. Reliability studies of *SCAT* have already been reported as part of the review of the test.

General Science Test (GST)

The content validity of the *GST* relies heavily on the procedures of the Examinations Branch, Department of Education, since the questions are taken from their bank. The questions in the bank originate with teachers who are contracted by the Branch to develop a given number of questions that meet listed specifications. These questions are then screened by the examination development officer and pretested in a number of randomly selected classrooms. On the basis of the pretest, the item is either placed in the bank or subjected to a revision and retested or discarded entirely. The criteria used for such judgments are: Difficulty between 0.20 and 0.80; Biserial Correlation over 0.30; Item Reliability Index of over 0.10. When an item is included in the bank it is classified on the basis of content and on cognitive level. The classification is the result of the judgment of a minimum of three people (the item author, the examination development officer, and one or more testing teachers). In addition, the classification is reviewed at the time an examination is put together from the bank. In the development of the *GST* three dimensions were considered important:

the cognitive level, the scientific process being tested and the subject area being used as an exemplar. The last is not important to the study but is a variable that had to be controlled.

The content validity with respect to the scientific process dimension of the test was defined in terms of a number of sources. The definition of those scientific processes that are to be measured are defined in Chapter I of this report. The processes that formed the criteria for item selection are inferring and hypothesizing.

The content and construct validity of the *GST* is in terms of the content of the items and how closely they measure the behaviors and the cognitive level described in the blueprint of the exam included as Table 3. Further, the construct validity is related to the behavioral descriptions of inferring and hypothesizing given by Tannenbaum (see Chapter II). The content validity can be assumed since the items are tested items taken from the item bank of the Department of Education. The construct validity of the test is in terms of the number of items that call upon students to:

Infer by:

1) identifying warranted conclusions from given observations;
2) identifying the important factors in a given set of circumstances;
3) matching an observation to a given conclusion;
4) differentiating between a statement of fact and a conclusion; and
5) recognizing alternative inferences about given observations.

Hypothesize by:

1) grouping conclusions into general statements about a phenomenon;

2) distinguishing between a proposition and a statement of fact about an observation;

3) identifying an important inference that supports a hypothesis; and

4) testing a hypothesis by identifying an experiment.

These behaviors, derived from Tannenbaum, were used as some of the criteria for selecting items from the bank. On the basis of this evidence a measure of construct validity is assumed. Further evidence is presented in Chapter IV as part of the factor analysis.

Another facet of construct validity is the extent to which the items measure the *higher mental processes and skills* identified by the Department of Education and reported earlier. The behaviors listed as being part of that cognitive level are related to the level of propositional thinking as defined in Chapter II.

The question of reliability hinges to a large extent upon the Kuder-Richardson reliability coefficient ($K\text{-}R_{20}$). In the pilot trial, the $K\text{-}R_{20}$ was 0.83 with an N of 290 students. In the subsequent administration the $K\text{-}R_{20}$ was 0.80 with an N of 801 students. In both cases the coefficient is quite sufficient to indicate a satisfactory degree of internal consistency.

Inference Test (IT) and *Hypothesis Test (HT)*

This study is concerned to a large extent with the question of the validity of these two tests. The question that comes to mind is:

What do these tests tell us about an individual's scientific process skills, mental process skills and propositional thinking ability? In answering this question we are concerned with the construct validity of the tests. The tests are valid to the extent that they sample an individual's inferring and hypothesizing skills.

It has been postulated in this study that hypothesizing is a cognitive skill related to what Piaget (1964) has called formal operations — a stage which many junior high students are in or are approaching. Further, since *IT* and *HT* are science content-free then any significant correlations with *GST* is evidence for construct validity.

The *construct validity* of the *IT* is closely related to the skills that the items elicit from responses. To determine this we must return to the behavior described by Tannenbaum (1971) and cited in Chapter II, that is that students should be able to demonstrate inferring by being able to

a) draw warranted conclusions from observations

b) identify the important factors in a given set of circumstances

c) relate an observation to a given conclusion

d) differentiate between a statement of fact about an observation and a conclusion arising from the observation

e) recognize that more than one inference may be drawn from a given set of data (p. 135).

In each item the observations are presented visually, the student has to relate the ball pattern to a specific target, the size of the target must be determined by its effect on a ball or number of balls, the ball pattern due to a number of targets must be sorted, and

different alternatives are to be weighed before making a decision. Further evidence of construct validity is contained in the discussion of factor analysis.

The *IT* format closely follows that of the Micciche and Keany (1969) description with the "targets" being unlimited by the physical limitations of the Hypothesis Machine. The answer sheet format was designed and the instructions were formulated. The items and the instructions were then examined by a nine-member critique panel composed of a selection of science teachers in central Alberta to make judgments about an aspect of content validity, that is, the suitability and appropriateness of the test for the target population. The questionnaire is included as Appendix D. Panel interaction was eliminated by having each member consider the test separately. An item was considered valid if six of the nine panel members agreed that the item difficulty was reasonable, seemed likely to function well with the students in the junior high school age categories and appeared to be of interest to those students. To further help them form an opinion, a group of nine students from a non-participating school were asked to try the test. Their responses are included with the panel's responses and are presented in Table 4. On the basis of these responses it was determined that the test should be used with only minor modifications in the instructions. Only one panel member identified items that should be modified. The students were quite intrigued by the novelty of the test but were unable to identify specific problems.

An aspect of the content validity of the *HT* was judged by the same panel and under the same conditions that passed judgment on the *IT*.

TABLE 4

VALIDITY QUESTIONNAIRE RESULTS:

INFERENCE TEST

	Panel Responses		Student Responses	
	Yes	No	Yes	No
1. Instructions had enough information?	9	0	7	2
2. Were the items as a whole:				
a) too easy			-	
b) easy	3		-	
c) too hard			1	
d) about right	6		8	
3. Did you think that the test was:				
a) interesting	9		6	
b) uninteresting			3	
c) waste of time			-	

4. Please identify those items you feel are inappropriate and should be changed.

The criterion for acceptance was again placed at two-thirds of the respondents.

The responses from the panel members are presented in Table 5. Three of the more complex problems were slightly modified and three letters that resulted in similar ball patterns were deleted as a result of the panel's response. The final version of the test has 20 targets and is included as Appendix B.

Data Processing

SCAT

The *SCAT* was processed in accordance with the normal scoring program of the Examinations Branch, Department of Education. This returns three scores for each respondent: a verbal *(SCAT-V)*, quantitative *(SCAT-Q)* and total score *(SCAT-T)*. It also gives a percentile, z-score, t-score. The scores for the two subtests and the total score are useful for drawing comparisons and calculating correlations.

GST, *IT* and *HT*

The *GST*, *IT*, and the *HT* results were processed by computer programs developed by the Division of Educational Research Services (DERS) and their experimental program library (XDER) of the University of Alberta.

The *IT*, *HT*, and *GST* results were processed by the DEST 02 program of the DERS library. This program returned the means, standard deviations, Pearson product moment-correlations, Kuder-Richardson Formula 20 reliability coefficient, the t-scores for each correlation coefficient and the probabilities associated with each t-score (DERS, DEST 02,

TABLE 5

VALIDITY QUESTIONNAIRE RESULTS:

HYPOTHESIS TEST

	Teachers		Students	
	Yes	No	Yes	No
1. Instructions had enough information?	9	0	7	2
2. Were the items as a whole:				
a) too easy				
b) easy				
c) too hard	2		3	
d) about right	7		6	
3. Did you think that the test was:				
a) interesting	9		7	
b) uninteresting			1	
c) waste of time			1	
4. Please identify those items that you feel are inappropriate and should be changed.				

July, 1969).

The *IT* and *HT* results were also processed by the NONP 10 program. This program returned frequency matrices for specified pairs of variables tabulated and printed in the form of cross-classification tables. Operations can be performed on these matrices yielding percentages, measures of association and tests of significance (DERS, NONP 10, February, 1970). The program returned the number of students in each grade answering each item, correctly or incorrectly; the number of boys and girls answering each item, correctly or incorrectly; and the number of students in each age category from 11 to 15 years answering each item, correctly or incorrectly. For each classification table the Pearson S contingency coefficient, C, or the Phi coefficient, ϕ, was computed from the Chi square statistic and the probabilities associated with the null hypothesis were calculated.

The *GST* results were processed by the TEST 01 program from the DERS library. This program returned the test mean and variance, the criterion mean and variance, and the test-criterion correlation. It also computed the $K\text{-}R_{20}$ reliability coefficient and a number of other statistics of use in modifying a test. It also returned information related to each item on the test, such as the item difficulty and the biserial correlation.

The *IT*, *HT* and *GST* results were also processed by the FACT 01 programs from both the DERS and XDER libraries. These programs are factor analysis packages which are designed to carry out a principal components factor analysis from the raw data or the Pearson product-moment correlation matrix. The principal axes factors are first

determined and then Varimax, Quartimax and Equamax orthogonal rotations are automatically applied. In addition, the XDER version does an oblique Procrustes rotation in which the axes are free to turn independently to achieve a best fit.

Between-test Relationships

For the analysis of relationships between tests, a stepwise regression analysis was applied after the test results had been combined into a matrix of scores from all students who wrote all four tests. The combining program was a FORTRAN routine that was developed particularly for this study. The stepwise regression program was the MULR 01 from the DERS library.

The data from the battery of tests were processed by the DERS ANOV 15 program. This program carried out a standard one-way analysis of variance applying the fixed-effect model for unequal observations in each group. The purpose in using this particular program was to determine the significance of the difference in the performance of boys and girls in the age categories from 11 to 15 on each of the *IT*, *HT* and combined tests.

The method used in the one-way analysis of variance is to compute the variances of the separate groups for mean differences. The scores of all subjects are then combined into a total score. If the variance of the combined total group is approximately the same as the average variance of the original subgroups, then there is no significant difference between the means of the original subgroups. If the variance of the total group is considerably larger than the average variance of the subgroups then a significant mean difference exists between two

or more of the subgroups. The test of the mean difference is accomplished with the use of the *F* statistic. This statistic is the ratio between the systematic or *between* source of variance due to the independent variable and the unsystematic, error, or *within* source of variance due to the uncontrolled variables. To the extent that the systematic variance is less than or equal to the unsystematic or error variance, the researcher would be unable to claim that any real differences exist. It is only when the *F* ratio is sufficiently greater than 1.0 for a given number of degrees of freedom that the researcher can make a claim of significant difference.

The multiple comparison of test means can be tested by use of the Newman-Keuls procedure in which the means are ordered from largest to smallest and a table of differences compiled from the ordered means. The differences are then tested by comparison with the "*q* statistic which has a distribution approximated by the 'studentized range distribution' having the parameters k = number of treatments and f = degrees of freedom for MS_{error}. The symbol $q_{0.99}$ (k,f) designates the 99th percentile point on the distribution" (Winer, 1962, p. 77). The *q* statistic is computed from

$$q_n = \frac{T_{largest} - T_{smallest}}{\sqrt{n \; MS_{error}}} \quad \text{(Winer, 1962, p. 77)}$$

where n is the number of observations, $(T_{largest} - T_{smallest})$ is the difference between two means, and MS_{error} is the mean-square experimental error. The Newman-Keuls procedure is a less powerful means of making a comparison but if differences exist then they will more likely be identified and further tests can be made of the significance of the

difference.

Results from those students who completed the entire battery of tests were retained for final analysis. In retaining only those subjects who completed the battery there is the chance that some of the information will be lost. This resulted in a reduction of the N to 539 students. The effect on the test statistics is illustrated by the data in Table 6.

The F statistic was computed as the ratio of the greater variance (S_g^2) to the lesser variance (S_l^2) for each of grade, age, etc., using the formula:

$$F = \frac{S_g^2}{S_l^2} \quad \text{(Popham and Sirotnik, 1967, p. 139).}$$

The t statistic was computed using one of the following formulae:

1. Separate variance model:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}}}$$

2. Pooled variance model:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{(n_1-1)\ S_1^2 + (n_2-1)\ S_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Selection of the appropriate t-test was made on the following basis:

when $n_1 = n_2$ and $S_1^2 = S_2^2$ the separate or pooled variance model is used with $n_1 + n_2 - 2$ degrees of freedom.

when $n_1 \neq n_2$ and $S_1^2 = S_2^2$ the pooled variance is used with $n_1 + n_2 - 2$ degrees of freedom.

when $n_1 = n_2$ and $S_1^2 \neq S_2^2$ the separate or pooled variance model is used with $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

TABLE 6

CHANGE IN TEST STATISTICS DUE TO ELIMINATION OF PARTIAL RESPONSES

	Restricted Group			Total Group				
	N	*(S)*	Mean	*N*	*(S)*	Mean	*F*	*t*
Grade	539	1.21	7.87	909	1.13	7.87	1.13	0
Age	539	0.99	13.67	909	1.00	13.69	1.02	0
Sex	539	0.55	1.53	909	0.50	1.53	1.00	0
SCAT-V	539	11.66	34.94	909	11.86	34.39	1.03	0.862
SCAT-Q	539	7.96	24.02	909	8.10	23.67	1.04	0.600
SCAT-T	539	17.76	58.96	909	18.16	58.06	1.05	0.924
IT	539	8.82	27.62	874	9.08	29.11	1.06	0.305
HT	539	38.84	64.19	887	37.73	59.95	1.06	2.16
GST	539	6.07	18.60	801	7.47	21.03	1.51	1.19

> when $n_1 \neq n_2$ and $S_1{}^2 \neq S_2{}^2$ the separate variance model is used with t-value determined by averaging the t-values for n_1-1 and n_2-1 degrees of freedom (Popham and Sirotnik, 1967, pp. 141-142).

The effect of curtailing the sample size can be seen to be minimal. The F statistic to compare the difference in variance shows only the *GST* as having undergone a significant change at the 2% level. An examination of the means of the sample indicated that there was no significant change in the results. The pooled variance model was used to determine the t statistic for all but the science test where the separate variance model was used and the t-values for n_1-1 and n_2-1 were identified and averaged. In no case was the t statistic significant at the 2% level for the two-tailed test. The pooled variance model results in a t-value which uses a greater number of degrees of freedom than is used with the separate variance model. Since a smaller t-value is needed to reject a given null hypothesis when a greater number of degrees of freedom are present, this indicates that the same t-value, when computed by the pooled variance formula, will be more likely to be significant than if it had been obtained by the separate variance formula. It follows that the pooled variance model results in a more powerful test, that is, one more apt to reject a null hypothesis. In this case the null hypothesis of *no difference* is not rejected in each case at the 2% level. The loss in confidence of the data appears to be minimal.

To obtain information relevant to the predictive validity of the *IT*, *HT* and *GST*, the test results from the 539 subjects were processed by the "Stepwise Regression" program MULR 06 of the DERS library. This program calculates a stepwise regression using the method of determinants

as described in Draper and Smith (1966, pp. 178-194). The program returns the test means, standard deviations, correlations of all variables, correlations of criterion and predictors, regression analysis of variance, F ratio and probability level, percentage of variance accounted for and regression weights for each variable entering at the specified level of significance, and the standard error of predicted y.

The program uses the stepwise procedure of M.A. Efroymson as described by Draper and Smith (1966, p. 78). Briefly, this method makes use of relationships that exist between variables to predict a criterion variable. As in most multiple regression procedures, stepwise regression enables one to arrive at a "best fit" prediction of the form:

$$y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

in which y is the dependent (criterion) variable,

X_1, X_2 ... are the independent (predictor) variables

b_0, b_1 ... are the coefficients that produce the "best fit," and

e is the error term (difference between the predicted and actual values of the dependent variable). The "best fit" is defined by the set of coefficients, b_0, b_1, ..., that makes the sum of e^2 a minimum for a particular series of criterion values and predictor values from a given sample. The stepwise regression procedure produces a series of intermediate regression equations of the form:

$$y = b_0^{(1)} + b_1 X_1^{(1)} + e^{(1)}$$

$$y = b_0^{(2)} + b_1^{(2)} X_1 + b_2^{(2)} X_2 + e^{(2)}$$

$$y = b_0^{(3)} + b_1^{(3)} X_1 + b_2^{(3)} X_2 + b_3^{(3)} X_3 + e^{(3)}$$

$$\begin{array}{cccccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}$$

in which the variable added at each step is the one which makes the greatest improvement in the "goodness of fit." In other words, the added variable accounts for the greatest proportion of the remaining variance of the dependent variable and also produces the greatest reduction in the sum of the standard error term. New coefficients are determined at each step to produce the "best fit" in terms of the specific variables included in the prediction equation. The reader is referred to the documentation for the program (DERS, MULR 06, 1969) and to Draper and Smith (1966) for a more complete exposition of the computation procedures used.

The data from the present study were analyzed by setting the probability level at .99 to ensure that all contributing predictor variables would be added and no predictors were dropped even when they ceased to contribute significantly. This procedure was followed to determine the effects, if any, of any suppressor variables.

CHAPTER IV

ANALYSIS OF DATA AND DISCUSSION

In this chapter the results of the investigation are presented in two parts. The first section contains the findings with respect to the validation of the instruments. The second part of the chapter contains the findings from the total battery of tests given to 539 students ranging in age from 11 to 15. Each of the hypotheses posed in Chapter I is tested.

Analysis of the *Inference Test*

Test Statistics

As a step in the establishment of the *IT* reliability, content and construct validity, the responses from 874 students were processed by computer programs from the DERS library, DEST 02 and NONP 10. The DEST 02 program calculated the means, variance, Pearson product-moment correlations, Kuder-Richardson formula 20 coefficient (K-R_{20}), the t values and the probabilities associated with each t. The program calculates the K-R_{20} reliability coefficient by using the variance-co-variance matrix.

As indicated in Table 7, the K-R_{20} reliability coefficient is 0.95 which would indicate that the 36 items tend to measure the same attribute since the coefficient is a measure of the internal consistency of the items. The means of the item responses varied from a high of 0.99 for item 1 to 0.50 for item 36, with an overall difficulty level of 0.78.

TABLE 7

MEANS AND VARIANCES FOR *INFERENCE TEST*

	Mean	Variance		Mean	Variance
Grade	7.87	1.63	Item 18	0.72	0.20
Sex	1.53	0.25	19	0.82	0.15
Age	13.67	0.98	20	0.79	0.17
Item 1	0.98	0.01	21	0.63	0.23
2	0.97	0.03	22	0.67	0.22
3	0.98	0.02	23	0.74	0.19
4	0.96	0.04	24	0.52	0.25
5	0.91	0.08	25	0.78	0.17
6	0.91	0.08	26	0.78	0.17
7	0.91	0.09	27	0.78	0.17
8	0.89	0.10	28	0.74	0.19
9	0.85	0.13	29	0.78	0.17
10	0.89	0.10	30	0.73	0.19
11	0.88	0.11	31	0.69	0.22
12	0.66	0.22	32	0.61	0.24
13	0.86	0.12	33	0.53	0.25
14	0.86	0.12	34	0.71	0.21
15	0.81	0.15	35	0.50	0.25
16	0.91	0.08	36	0.50	0.25
17	0.86	0.12			

$KR_{20} = 0.952$

N = 874 (408 boys, 466 girls)

Grade 7 - 272
Grade 8 - 326
Grade 9 - 276

This level of difficulty and the low variances reported lends credence to the validity of the $K\text{-}R_{20}$. The variances ranged from 0.01 to 0.25 for the item responses.

Item Correlations

The correlation matrix shows a similar pattern to that obtained from the factor analysis and so is not reported. The probability that there is a correlation between items was tested with t values calculated from the correlation matrix using the formula:

$$t = r_{ij} \frac{N-2}{\sqrt{1-r_{ij}}} .$$

The null hypothesis that *there is no correlation between items* was tested. With 873 degrees of freedom at the 5% level of significance the correlations must exceed 0.075 (Popham and Sirotnik, 1973, p. 387).

The correlation matrix produced from the *IT* scores indicates some patterning among items. This is on the basis of shape, number of targets, that is, the same level of complexity. These patterns among items with 0.500 are depicted in Figure 1. The asterisks on the grid show some patterns among simple items (1 element), moderately complex (2 elements), and complex (3 or more elements).

The program produced t-scores, and the probability level associated with each t-score, for each correlation in the matrix. These are not reproduced because of the lack of space. Item 1 was the only item to have statistically insignificant correlations with other items (16, 17, 20, 24, 26, 27, 29, 30, 33, 35 and 36). This is probably because item 1 was the only one without any hidden target and was an obvious pattern

FIGURE 1

INTER-ITEM CORRELATIONS FOR THE *INFERENCE TEST*

Item		4	[illegible]	9	10	11	14	20	23	25	26	27	28	29	30	31	32	33	34	36
	No. of elements involved	(1)	(1)	(1)	(1)	(1)	(2)	(2)	(3)	(2)	(2)	(2)	(2)	(2)	(4)	(2)	(4)	(4)	(3)	(6)
3	(1)	*																		
6	(1)		*																	
8	(1)			*	*	*	*	*				*	*							
9	(1)				*	*														
10	(1)					*	*													
11	(1)						*													
14	(2)							*				*		*						
15	(2)							*												
20	(2)								*	*	*	*	*	*	*	*			*	
22	(3)								*		*									
23	(3)									*	*	*	*	*	*	*	*		*	
25	(2)										*	*	*	*	*	*	*		*	
26	(2)											*	*	*	*	*	*		*	
27	(2)												*	*	*	*	*		*	
28	(2)													*	*	*	*		*	*
29	(2)														*	*	*		*	
30	(4)															*	*	*	*	
31	(2)																*	*	*	
32	(4)																	*	*	*
35	(6)																			*

*Indicates correlation greater than 0.500.

for most students. All of the other item correlations were statistically significant at the 5% level and most were significant beyond the 1% level.

The addition of student variables adds to the information about the relationships that exist between items. Statistically significant correlations between items and student variables are shown in Table 8. The most important student characteristic is the grade level, which correlates with 31 items. The next important item is the school which correlates significantly with 28 items, followed by age (20 items) and sex (17 items).

Cross-classification

The data from this test were also processed by means of the NONP 10 program of the Division of Educational Research Services. This program gives frequency distributions for specified pairs of variables that are reported in the form of cross-classification tables. Operations were performed upon these matrices yielding percentages, measures of association, and tests of significance (DERS, NONP 10, February 1970).

The program returned the number of students in each grade answering each item correctly or incorrectly; the number of boys and girls answering each item correctly or incorrectly; and the number of students in each age category from 11 to 15 years answering each item correctly or incorrectly. For each classification table, the Pearson S contingency coefficient, C, or the Phi coefficient, ϕ, was computed from the Chi square, χ^2, statistic and the calculated probabilities associated with the null hypothesis from the computer program: *that there is no relation between each of the items and age, grade or sex.*

TABLE 8

CORRELATIONS BETWEEN STUDENT VARIABLES AND

RESPONSES ON *INFERENCE TEST*

Student Variables	Items Correlating at 5% or Better
School	5,6,7,8,9,10,12,13,14,15,17 to 34 inclusive
Grade	2,3,4,7,8,9,10,12,13,14,15,17 to 36 inclusive
Sex	2,8,9,10,11,12,14,15,18,20,23,25,26,27,29,30 and 33
Age	15,16,19 to 36 inclusive

Of the 874 students responding to the test, most responded to all of the items. Even the most difficult item, number 36, had 714 responses which represents about 82% of the total students writing the test. This supports the assumption that the test was not highly speeded and that the $K\text{-}R_{20}$ coefficient of 0.845 is not spuriously high. The obtained value of the $K\text{-}R_{20}$ coefficient then can be taken as support for the large number of significant correlations among the items. These correlations are also taken as evidence that there is a significant construct validity.

The cross-classification tables are not included because of space limitations but those items with significant χ^2 are identified in Table 9. A large number of items do associate significantly with grade level, sex and age. Comparing Table 9 with Table 8, it should be noted that the 17 items with significant χ^2 statistics also have significant correlations with grade level. It may be concluded that the test results are positively related to a student's grade level.

The classification tables for sex and item success resulted in contingency coefficients which ranged from -0.01 to 0.14 with 22 items having χ^2 values beyond the critical value for significance at the 5% level. This means that there is a positive association between item success and the sex of a student on 22 of the 36 items. In comparing Table 9 with Table 8 the items that appear related to sex are very similar. Only five items appear in that category in Table 9 that do not appear in Table 8.

The classification tables for age categories 11 to 15 with item success resulted in contingency coefficients from 0.05 to 0.24 with

TABLE 9

χ^2 BETWEEN STUDENT VARIABLES AND ITEM RESPONSES

ON *INFERENCE TEST*

Student Variables	Items with Significant χ^2 (5% or better)
Grade	9,12,13,15,19,20,21,23,24,25,27,30 to 35
Sex	2,6 to 12,14,15,17,18,20,22,23,25 to 30, 32,33
Age	1,12,18,20,21,27 to 32, 34 to 36.

14 of the χ^2 values beyond the critical value for significance at the 5% level. Again comparing Table 9 with Table 8, it is evident that only three items that have significant χ^2 do not also have significant correlations and only three items that have significant correlations do not have significant χ^2 values. It is evident that some items are more difficult for younger students than for older. There appears to be increased competence on the *IT* with age.

Factor Analysis

The item data from the *IT* were processed by the DERS, FACT 01, factor analysis package which is designed to carry out a principal components factor analysis from the raw data or the Pearson correlation matrix. The principal axes factors were first determined and then Varimax, Quartimax and Equamax orthogonal rotations were automatically applied. The student variables were then added to the data matrix and processed.

This analysis was undertaken to provide a simplified description of interrelationships among the items and the independent variables, in other words, to determine the simple structure of relationships between the test items and between the test items and certain student variables.

The first factor analysis resulted in an unrotated factor matrix that showed factor loadings on one dimension that accounted for almost 73% of the total variance of the test and 84% of the common variance. This supports the assumption that the items are all measuring the same skill. On rotation, the variance was distributed among the four factors that had eigenvalues greater than 1.0. The factor loadings tend to

cluster according to the number of different shapes and complexity of the problem presented by the item. For example in Table 10, factor 1 has loadings from items that present a number of targets of the same shape and a number of targets of different shapes in combination. On the other hand, factor 2 has loadings from the simpler targets involving single shapes or simpler combinations. Factor 3 has loadings from moderately difficult problems and factor 4 has loadings from items that have a specific target (i.e. a cup) alone or in combination.

The addition of respondee variables resulted in a slightly different loading pattern. The unrotated factor matrix showed factor loadings on one dimension that account for 35% of the total variance and 58% of the common variance. Further, the addition of personal data spread the variance among 8 factors with eigenvalues greater than 1.0 and reduced the accounted variance to 61.4%. This loading primarily on a single factor is further evidence of the uniformity of the items in calling on a single dimension of the skills of the respondees.

The rotation of the axes in a Varimax solution resulted in the loadings being maximized on the eight factors which are presented in Table 11. Six of the factors have items loading in a pattern similar to that obtained without the addition of the respondee variables.

The addition of student variables has added four factors and has changed the definitions of the factors so that:

Factor 1 has loadings from items that reflect the most difficult combinations that are placed towards the end of the test.

Factor 2 has loadings from items that use a slope as the hidden target.

TABLE 10

EQUAMAX ROTATED FACTOR MATRIX FOR *INFERENCE TEST*

Item	Factors 1	2	3	4	Communalities
1		95*			104
2	36	62*	50	48	99
3		64*	64*	33	98
4	36	64*	52	50	105
5		67*		58	84
6		66*		68*	93
7		67*		63*	91
8	63*	51	44	32	94
9	48	67*	42		87
10	58	64*	37		94
11	54	59	41		85
12	42	73*	37		84
13		57	30	58	74
14	48	46	52	31	81
15	43	34	46	48	74
16	37		46	53	82
17			36	76*	81
18		50	63*	36	80
19	31		43	69*	79
20	55		59	41	89
21		38	69*	32	74
22		34	70*	34	78

(Cont'd)

TABLE 10 (Cont'd)

Item	Factors 1	2	3	4	Communalities
23	50	31	54	44	83
24	33		60*	37	69
25	61*		57	37	86
26	72*		53	40	98
27	75*		56	37	103
28	75*		47	31	93
29	70*		52	37	94
30	68*		55	33	93
31	64*		41	43	84
32	72*	32	31	36	84
33	53		45	42	72
34	66*	31	34	43	84
35	74*			46	80
36	73*	36		39	84
% Total Variance	25.1%	21.4%	21.2%	19.2%	Sum. 31.14

Total variance accounted for 86.9%.

*Items that serve to define the factor.

Note: The entries in the above matrix have been multiplied by a factor of 100 and rounded to the nearest whole number. Loadings less than 30 have been dropped for simplification.

TABLE 11

ROTATED FACTOR LOADING MATRIX: ITEM AND STUDENT VARIABLES

FROM *INFERENCE TEST*

	Factors							
Data	1	2	3	4	5	6	7	8
School								92
Grade							86	
Sex		-43						
Age							87	
Item 1					62			
2					64			
3					77			
4					72			
5			60					
6			80					
7			79					
8	37	60						
9	30	63						
10	30	68						
11	31	67						
12		51				35		
13			39	35				
14	34	54		34				
15	36	37		42				
16				66				
17			34	58				

(Cont'd)

TABLE 11 (Cont'd)

Data	Factors 1	2	3	4	5	6	7	8
Item 18						58		
19	37			51				
20	56	35		37				
21	30					63		
22	38					56		
23	60					34		
24	42					55		
25	69			30				
26	76							
27	80							
28	78							
29	73			30				
30	78							
31	74							
32	73							
33	59					40		
34	76							
35	61							
36	63							
Variance:	8.49	3.40	2.64	2.52	2.45	2.39	1.60	1.06
% of Total Var.	21.24	9.51	6.59	6.29	6.13	5.98	4.00	2.64

Total variance accounted for: 61.37%; Sum of communalities: 24.55

Note: The entries in the above matrix have been multiplied by a factor of 100 and rounded to the nearest whole number. Loadings of less than 30 have been dropped for simplification.

Factor 3 has loadings from items that have a cup as a hidden target.

Factor 4 has loadings from items that have a moderately difficult pattern of shapes.

Factor 5 has loadings from items that have a block as a hidden target.

Factor 6 has loadings from items that use cup-slope targets in combination.

Factor 7 has loadings from grade and age with only minor loadings from the items and seems to be an experience factor which loads only lightly on the item responses.

Factor 8 has loadings from the school code and this code seems highly correlated with differences in the *SCAT* scores so it may be a scholastic ability factor which loads very lightly on the item responses.

The first six factors are related to differences in the complexity and shape of the targets. Factors 1 and 4 seem to have loadings from groups of items that involve a number of elements of differing shapes and sizes (number of columns affected). It is proposed then that there is a difference in the level of inferring ability that is related to the number of parts of a question that must be considered before arriving at a conclusion. On a more physical level there is a qualitative difference in an inference that is made from a single event or fact as opposed to one made from an observation of a related series of events or facts.

Summary of Analysis of the Inference Test

It is evident from these analyses of the results from the *IT* that the test has a respectable reliability coefficient (0.952) and that most students have the capacity to complete the test (difficulty of 0.78). These two test characteristics are related to content validity in that they establish limits on the validity of a test. A test with stability and within the capacity of the subjects to answer (its answerability) does not necessarily have content validity, but if it is not stable and is "unanswerable" it cannot have content validity. Further evidence of content validity was presented in Chapter III, namely, that in the judgment of a panel of teachers and students the test was suitable and measured the described behaviors.

The construct validity of the *IT* is indicated by two pieces of evidence, the pattern of item correlations presented in Figure 1 and the factor structure presented in Tables 10 and 11. The simple structure indicates one main factor, with loadings from 24 items in Table 11 and 26 items in Table 10 apparently linked to an ability to make inferences from information that could lead to more than one conclusion. Factors 2, 3 and 5 seem more closely related to the shape of the hidden target with number 5 being the most readily identifiable target of the three used. Factors 4 and 6 are related to the ability to infer combinations of unseen targets of moderate complexity. The simple structure breaks down into two main components, one being related to the test items and their combination of targets and the other being related to the student's ability to perceive the unseen target. It is this second cluster of related factors (1, 4 and 6) that provides some evidence for construct

validity of the *IT*. The main argument for construct validity is the simplicity of the factors identified in Tables 10 and 11. Part of the definition of construct validity presented in Chapter II is that a test that has construct validity has a relatively simple factor pattern.

Analysis of the *Hypothesis Test*

Test Statistics

In establishing the reliability, content and construct validity of the *HT*, the responses from 887 students were processed by the DERS computer programs, DEST 02, NONP 10, and FACT 01, in the same fashion described for the *IT*. As indicated in Table 12, the K-R_{20} reliability coefficient is 0.936 which would indicate that the 20 items in the test tend to measure the same attribute since the K-R_{20} is a measure of internal consistency and stability. The total test mean is 60.50, the standard deviation is 37.58 and the difficulty level is 0.303. As is indicated in Table 12, the variances of the items range from 4.11 to 12.85 and reflect a greater dispersion among the scores, which indicates that the *HT* is a more difficult test than the *IT*.

Item Correlations

The correlations between each of the items and the student data are shown in Table 13. The pattern indicates a closer relationship among items close together on the test. That is, there appears to be some relationship between the targets and how students responded to them. The effects of school, grade, sex and age appear to be minimal. There seem to be four or five affinity groups among items which indicates that

TABLE 12

MEANS AND VARIANCES FOR *HYPOTHESIS TEST*

	Mean	Variance
Grade	7.9	1.24
Sex	1.53	0.25
Age	13.67	1.02
Item 1	5.72	5.36
2	5.14	6.95
3	5.28	7.80
4	4.55	7.77
5	3.78	9.66
6	3.88	10.90
7	3.85	12.85
8	3.23	10.58
9	2.76	10.55
10	2.67	11.84
11	2.25	10.10
12	2.10	9.82
13	1.83	9.29
14	1.38	6.00
15	1.22	5.25
16	1.24	5.68
17	1.23	6.06
18	0.87	4.11
19	0.93	4.47
20	1.03	5.16

KR_{20} = 0.936

N = 887 (414 boys, 473 girls)

Grade 7 - 274
Grade 8 - 335
Grade 9 - 278

TABLE 13

CORRELATION MATRIX FOR THE *HYPOTHESIS TEST* ITEM RESPONSES

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.00																			
2	0.70	1.00																		
3	0.61	0.68	1.00																	
4	0.51	0.62	0.64	1.00																
5	*	0.56	0.53	0.60	1.00															
6	*	*	*	*	0.69	1.00														
7	*	*	*	*	0.55	0.77	1.00													
8	*	*	*	*	0.52	0.72	0.82	1.00												
9	*	*	*	*	*	0.61	0.65	0.71	1.00											
10	*	*	*	*	*	0.50	0.57	0.58	0.81	1.00										
11	*	*	*	*	*	*	0.55	0.56	0.75	0.85	1.00									
12	*	*	*	*	*	*	0.51	0.54	0.73	0.83	0.88	1.00								
13				*	*	*	*	*	0.54	0.64	0.66	0.71	1.00							
14	*	*	*	*	*	*	*	*	0.56	0.64	0.67	0.74	0.79	1.00						
15		*	*	*	*	*	*	*	0.54	0.62	0.65	0.71	0.77	0.89	1.00					
16		*	*	*	*	*	*	*	0.52	0.59	0.60	0.68	0.74	0.82	0.86	1.00				
17				*	*	*	*	*	*	0.53	0.54	0.61	0.70	0.73	0.75	0.84	1.00			
18	*	*		*	*	*	*	*	*	*	*	0.54	0.58	0.67	0.66	0.75	0.85	1.00		
19				*	*	*	*	*	*	0.50	0.51	0.58	0.63	0.69	0.71	0.79	0.89	0.88	1.00	
20						*	*	*	*	*	*	0.50	0.68	0.59	0.63	0.68	0.82	0.72	0.82	1.00

*Significant correlations less than 0.50

factor analysis should reveal a simple structure amenable to interpretation.

The computer program calculated the *t*-values for each correlation and the probability that the correlation was due to chance. On this basis, the non-significant correlations were dropped from the table and the significant but minor correlations have been indicated by asterisks in an attempt to reduce the size of the table.

Cross-classification

The data from this test were processed using the DERS library NONP 10 program in the same fashion as for the *IT*. This program returned frequency matrices for specified pairs of variables, tabulated and printed in the form of cross-classification tables. Operations can be performed upon these matrices yielding percentages, measures of association and tests of significance.

The cross-classification tables are not included because of space considerations but the χ^2 and its probability for each of the three criteria is included in Table 14.

As presented in Table 14, the χ^2 statistic for item response vs. grade ranges from 49.00 to 14.44. Only one item has a significant χ^2 with grade (3) at the 5% level or better. For the sex vs. item performance, the χ^2 ranges from 20.94 to 2.73 and five items have significant statistics (4, 11, 12, 15 and 16) at the 5% level or better. For age category vs. item performance the χ^2 ranges from 80.38 to 37.90 but only three items have significant statistics (5, 7 and 12) at the 5% level.

TABLE 14

χ^2 FOR THE *HYPOTHESIS TEST*

Item	Grade		Sex		Age	
	χ^2	p	χ^2	p	χ^2	p
1	28.18	0.25	4.17	0.84	55.25	0.22
2	29.01	0.22	9.57	0.30	41.35	0.74
3	49.00	*0.00*	13.05	0.11	46.98	0.51
4	23.67	0.48	15.97	*0.04*	61.03	0.10
5	35.69	0.06	8.04	0.42	12.74	*0.01*
6	32.44	0.11	2.91	0.94	60.24	0.11
7	24.05	0.46	4.39	0.82	74.36	*0.01*
8	24.43	0.44	7.19	0.52	48.41	0.46
9	30.17	0.18	5.18	0.74	57.83	0.16
10	27.41	0.29	12.03	0.15	46.42	0.42
11	34.44	0.08	18.29	*0.02*	54.65	0.24
12	35.01	0.07	18.51	*0.02*	80.38	*0.02*
13	16.65	0.86	11.57	0.17	53.44	0.08
14	23.87	0.47	4.17	0.84	48.00	0.18
15	26.05	0.35	20.94	*0.01*	37.90	0.57
16	25.35	0.39	19.34	*0.01*	39.73	0.48
17	34.45	0.08	10.45	0.23	45.54	0.25
18	12.89	0.91	5.86	0.56	41.72	0.20
19	21.96	0.58	2.73	0.95	44.88	0.27
20	14.44	0.94	2.94	0.94	41.95	0.39

We may conclude then that there appears to be a relationship between age, sex and grade with success on only eight items. Twelve of the items appear to be equally difficult for students no matter what their grade, age or sex. However, more evidence will be presented when the mean differences between groups are examined.

Factor Analysis

The item response data from the *HT* and the student variables were processed by the DERS FACT 01 program, which is a factor analysis package designed to carry out a principal components factor analysis from the Pearson correlations. The principal axis factors are first determined and then Varimax, Quartimax and Equamax orthogonal rotations are applied. The Varimax rotated loading matrix is presented in Table 15. The factor structure shows the test to be essentially a three-factor test.

The unrotated matrix shows loadings for items 9 to 20 to be on the first factor, with items 5 to 15 loadings on factor 2 and items 1 to 8 loadings on factor 3. Factors 4 and 5 include loadings from the grade level, age category and sex. Factor 1 has salient loadings from items 13 to 20 which were the most difficult items according to the item analysis. Factor 2 has salient loadings from items 6 to 12 which were moderately difficult. Factor 3 has loadings from items 1 to 5 which were the easiest items. Factor 4 appears to be an experiential factor since the grade and age information received the largest loadings. The fifth factor appears to be a sex-linked and scholastic factor which also contributes little to the item variance. The three levels

TABLE 15

VARIMAX FACTOR MATRIX FOR THE *HYPOTHESIS TEST*

	Factor				
	1	2	3	4	5
School					46
Grade				88	
Sex					89
Age				87	
Item 1			82		
2			86		
3			82		
4			76		
5		46	67		
6		70	48		
7		75	38		
8		77	32		
9	35	79			
10	47	75			
11	50	73			
12	58	67			
13	75	40			
14	78	40			
15	81	37			
16	87				
17	93				
18	88				
19	91				
20	86				
Variance	6.80	4.72	3.67	1.56	1.06
% of Total Var.	28.35	19.66	15.31	6.52	4.42
% of Common Var.	38.17	26.47	20.61	8.78	5.96

Total of variance accounted for = 74.288%; Sum of communalities = 17.829

Note: The entries in this table have been multiplied by a factor of 100 and rounded off to the nearest whole number. Loadings less than 30 have been dropped for simplification.

of difficulty appear to account for 63.4% of the total variance of the test. The respondee variables account for only about 11% of the total variance. This confirms that success on the test is relatively independent of the age, grade and sex of the individual.

Summary of the Analysis of the Hypothesis Test

It is evident from the analysis of the data from the *HT* that the test has a respectably high reliability (0.936), but as it is quite difficult for many students the K-R_{20} may be optimistically high. However, the factor analysis showed a very simple three-factor structure with high loadings from the items on only one factor. This would indicate a high level of internal consistency. The test difficulty is to be expected since it has been developed to measure a skill that many students have either newly acquired or have not yet developed. The content validity will depend upon a number of pieces of indirect evidence. One is that the panel of judges felt that the test was suitable and in their judgment called for the use of the described behaviors. The other piece of evidence is that the correlation pattern and factor pattern indicate a high degree of internal consistency on the part of the items. The identification of the particular attribute being measured will be discussed under concurrent validity. Suffice it to say that the *HT* is measuring a cognitive skill that many students have found difficult to exhibit.

Analysis of the
General Science Test

Test Statistics

Returns were obtained from 801 students and these data were processed by the University of Alberta Division of Educational Research Services computer programs TEST 01, DEST 02, and FACT 01.

The first program, TEST 01, returned the test mean and variance, the Kuder-Richardson formula 20 reliability coefficient, and the test difficulty level. These test statistics are:

Test mean	-	21.03
Test variance	-	55.84
K-R_{20} reliability	-	0.818
Test difficulty	-	0.421.

This difficulty level is not unexpected since the choice of items was based on a higher than usual criterion of mental skill. The item statistics indicated in Table 16 show a satisfactory performance in terms of the purpose for which the items were developed. The difficulty level ranges from 0.20 to 0.70 and the biserial correlation ranges from 0.20 to 0.59, both within the defined limits.

The second program used, DEST 02, computed the means, standard deviations, Pearson product-moment correlation, t-score values and the probabilities associated with each t-value for each item, and the Kuder-Richardson formula 20 coefficient for the test by using the variance-covariance matrix. The 50 x 50 matrix is not reported because of space limitations.

TABLE 16

ITEM STATISTICS FOR *GENERAL SCIENCE TEST*

Item	Difficulty	Biserial Correlation	Item	Difficulty	Biserial Correlation
1	0.46	0.51	26	0.59	0.59
2	0.42	0.45	27	0.42	0.46
3	0.20	0.52	28	0.20	0.43
4	0.33	0.39	29	0.28	0.30
5	0.41	0.47	30	0.57	0.53
6	0.50	0.37	31	0.70	0.50
7	0.37	0.36	32	0.55	0.51
8	0.38	0.45	33	0.51	0.44
9	0.28	0.26	34	0.20	0.20
10	0.46	0.43	35	0.53	0.52
11	0.33	0.39	36	0.46	0.41
12	0.45	0.48	37	0.56	0.52
13	0.27	0.36	38	0.23	0.24
14	0.71	0.50	39	0.60	0.42
15	0.38	0.25	40	0.32	0.38
16	0.40	0.29	41	0.36	0.57
17	0.26	0.36	42	0.42	0.36
18	0.42	0.25	43	0.39	0.27
19	0.27	0.29	44	0.62	0.49
20	0.24	0.31	45	0.59	0.53
21	0.51	0.58	46	0.38	0.40
22	0.39	0.33	47	0.40	0.45
23	0.72	0.52	48	0.29	0.38
24	0.62	0.37	49	0.43	0.25
25	0.29	0.30	50	0.35	0.35

Factor Analysis

The correlation matrix showed very little patterning with most correlations significant at the 5% level or better. To check on the interrelationships, the data were processed by means of FACT 01.

The FACT 01 program is the factor analysis package of the Division of Educational Research Services XDER library developed in August 1974. This program carries out a principal components factor analysis from either raw data or a correlation matrix. Varimax, Quartimax, Equamax orthogonal rotations are applied to the principal axes factors.

The program returned 19 factors with eigenvalues greater than 1.0. These with the corresponding item loadings and the percentage of the total test variance that is attached to each factor are listed in Table 17.

Examination of the item loadings on each factor and the plot of each pair of factors suggests that the following interpretations can be made of the factors:

Factor 1 - the ability to relate experimental observations to everyday examples.

Factor 2 - the ability to infer a cause and effect relationship between two observations.

Factor 3 - the ability to formulate hypotheses about observations in terms of previously learned physical laws.

Factor 4 - the ability to relate hypotheses to observations.

Factor 5 - the ability to infer explanations from observations.

Factor 6 - the ability to apply hypotheses to real situations.

Factor 7 - the ability to identify experimental variables which

TABLE 17

VARIMAX FACTOR MATRIX: FOR *GENERAL SCIENCE TEST*

Factor	Item Number/Loading								% of Total Variance
1	21/39	23/63	24/77	26/50	27/52	30/41	31/32	35/39	3.70
2	5/78	14/50	21/42	31/30	35/34	37/33	47/45		3.55
3	3/33	8/70	10/43	12/57	36/30	37/38	47/34	50/42	3.52
4	37/34	41/43	42/35	44/78	45/73				3.40
5	14/33	28/80	29/83						3.37
6	19/-32	30/40	36/30	37/30	39/66	42/55			3.30
7	3/36	10/37	11/80	20/37	36/50				2.87
8	4/57	43/58	46/54	50/51					2.82
9	1/36	7/70	40/32						2.77
10	3/31	17/51	22/67	27/32	33/39				2.71
11	48/74	49/82							2.68
12	1/33	13/37	14/34	23/-31	25/80				2.57
13	19/38	31/-34	34/83						2.44
14	6/65	20/-31	41/41						2.38
15	2/41	38/81							2.37
16	9/81	10/-38							2.36
17	18/83	19/45							2.35
18	1/36	15/75	40/39						2.35
19	13/-47	16/81							2.20

Total variance accounted for - 53.71; Sum of communalities - 27.38

Note: The loadings in the above table have been multiplied by a factor of 100. Loadings less than 30 have not been reported for simplification.

can be controlled.

Factor 8 - the ability to discern practical applications of physical attributes.

Factor 9 - the ability to combine a number of inferences into a hypothesis.

Factor 10 - the ability to infer causal relationships.

Factor 11 - the ability to identify a correct inference about an experimental situation.

Factor 12 - the ability to distinguish observations from inferences.

Factor 13 - the ability to relate observations to hypotheses.

Factor 14 - the ability to relate observations to specific hypotheses.

Factor 15 - the ability to interpret observations in terms of hypotheses.

Factor 16 - the ability to relate simple hypotheses to practical applications.

Factor 17 - the ability to develop an inference into a more general hypothesis.

Factor 18 - the ability to relate an observation to the best hypothesis.

Factor 19 - the ability to identify important assumptions and variables in an experimental setting.

The 19 factors can be grouped into the broad categories that were originally used to design the test, that is, those scientific process skills that were part of the identifying criteria for choosing items.

Inferring: Factors that have loadings from items that deal with

the generation of inferences, the relating of an inference to a hypothesis and the distinguishing of an inference from an observation. These are factors 2, 5, 10, 11, and 12.

Hypothesizing: Factors that have loadings from items that deal with the formulation of hypotheses, the distinguishing of a hypothesis from an inference, the relating of an observation to a hypothesis and the use of established hypotheses to explain new information. These factors are: 3, 4, 6, 9, 13, 14, 15, 16, 17, and 18.

Other scientific processes: Other factors are those relating to the identification of variables and assumptions (7, 16 and 19), application of hypotheses, inferences and general principles to new situations (6, 8 and 16).

Summary of the Analysis of the General Science Test

The factor analysis has served as another link in establishing the construct validity of the *GST*. There is little doubt that the test is measuring mental processes that are either directly or indirectly linked with the making of inferences and formulating of hypotheses. It is also clear from the factor analysis that the questions are multidimensional in that they involve a variety of facts, reasoning skills and mental abilities. This is to be expected in a bank of questions that have been developed to measure the attainment of both higher levels of cognitive functioning and science skills. Their inclusion in the battery of tests is necessary, however, in establishing the concurrent and predictive validity of the *IT* and the *HT*.

The test proved to be quite difficult for the younger students who had not developed these skills at the requisite level. The $K\text{-}R_{20}$

coefficient is quite high in spite of the rather complex "simple structure" that evolved from the factor analysis. This level of $K\text{-}R_{20}$ is probably related to both the internal consistency of the test as indicated by the item statistics and the overall difficulty level, both of which would tend to cause the test to have a high reliability. The question is: which one has had the greatest effect? In the absence of other evidence, the assumption is made that internal consistency has had the greatest effect.

Analysis of the *Cooperative School and College Ability Test*

The raw scores of the *SCAT* were processed by the Department of Education *SCAT* program. This program returned frequency listing, percentile, *z*-score, *t*-score, cumulative percentage mean, variance and standard deviation of the verbal, quantitative and total scores for the total group of students, and for each school. From the means and variances the Kuder-Richardson reliability coefficient was computed, using formula 21 for the total group scores on the verbal, quantitative and total tests. The test statistics were then used by the program to produce scaled scores for each student and these scores were then included as part of the total test scores in the final analysis.

The test statistics are presented in Table 18. In the tables the subscores and the total score will be referred to as *SCAT*-V for the verbal subtest, *SCAT*-Q for the quantitative subtest and *SCAT*-T for the total score.

The reliability estimate is reasonably close to that obtained in other administrations of the test in Alberta. The means and variances

TABLE 18

TEST STATISTICS FOR THE

COOPERATIVE SCHOOL AND COLLEGE ABILITY TEST

	Mean	Variance	Standard Deviation	$K\text{-}R_{21}$
Verbal test	34.39	104.77	11.86	0.76
Quantitative test	23.67	65.58	8.10	0.65
Total Test	58.06	329.69	18.16	0.83

N = 1201

are different than those obtained in the other administrations but this is to be expected since the test is usually administered only to grade 9 students while in this study grades 7 and 8 students were included in the sample.

The *SCAT* data from each of the 11 schools in the study were compared to see if the students varied significantly in scholastic ability. These comparisons are presented in Table 19.

In the first step in the comparison the assumption of homogeneity of the variance was tested by comparing the lowest and highest variances of the subgroups. When this was done, the F_{max} test did not show significance at the 5% level. A further check showed that the extreme variances were not significantly different beyond the 5% level. On this basis the pair-wise comparisons were made. The observed differences and *z*-scores are presented in Table 19.

For this table, *z* was computed using the formula:

$$z = \frac{\bar{X}_T - \bar{X}_x}{\sqrt{\frac{S_T^2}{N_T} \quad \frac{S_x^2}{N_x}}}$$

For a two-tailed test at the 1% level of significance the *z*-score must exceed 2.576. Only three schools exceed this value. School 2 has a *z*-score of 5.76, school 9 has a *z*-score of 2.858, and school 11 has a *z*-score of 3.166. For a two-tailed test at the 5% level the *z*-score must equal or exceed 1.960, which adds only school 8 to

TABLE 19

MEANS, VARIANCES AND MEAN DIFFERENCES FOR THE

COOPERATIVE SCHOOL AND COLLEGE ABILITY TEST

School	*N*	Mean	Variance	Mean difference	*z*-score
1	80	55.88	243.13	-2.18	1.329
2	128	50.63	259.20	-7.43	5.760**
3	43	54.21	458.68	-3.85	1.199
4	240	59.44	336.84	+1.38	1.353
5	20	50.85	340.13	-7.21	1.767
6	109	61.94	307.60	+3.88	2.471*
7	83	60.69	353.61	+2.63	1.328
8	54	53.94	249.76	-4.12	2.000
9	15	48.40	177.04	-9.66	2.858**
10	83	56.36	290.26	-1.70	0.960
11	62	64.74	299.03	+6.68	3.166**
Combined:	917	58.06	329.69		

*Significant at 5%

**Significant at 1%

the ranks of schools that have significantly different *SCAT* results with a z-score of 2.471. The fact that four out of 11 schools showed a difference in the general level of scholastic ability would tend to make school a slight, indirect indicator of scholastic ability and would tend to explain the correlation of school with the test results. A partial correlation, holding *SCAT* constant, succeeded in changing the correlation of school with *IT* (0.153), *HT* (0.127) and *GST* (0.115). The change in the *GST* vs. school correlation is the greatest and is meaningless with respect to the study.

By "partialling out" the effect of *SCAT*, the influence of the school variable is reduced a little but not to the point of changing the significance of the correlations.

Concurrent Validity

Discussion

The concurrent validity of a test, according to conventional definitions, is the extent to which the test correlates with other tests measuring similar attributes. The three tests developed for this study are related to external referents to establish their validity. As each test becomes valid, it in turn adds to the validity of the others. That is, as the content, construct and concurrent validity of each test is strengthened, that test then can be used to validate the others. In the present stage of the analysis each test has been judged to measure the behaviors that defined the basic construction of the test and hence has content validity. Each test has been established as being reliable, that is, is internally

consistent. In addition, the factor analyses have shown each test to have a simple structure related to the defining constructs and therefore is said to have construct validity.

Correlations Among Tests

Table 20 presents the Pearson correlations among the four tests in the battery. It should be noted that the test of the null hypothesis, *that there is no correlation between the tests*, in each case has resulted in four of the correlations being found to be significant at the 1% level. Three of the four correlations are not particularly large but they do indicate important relationships among the *GST*, *IT* and *SCAT*. This relationship is interpreted as indicating a similarity in the abilities being measured. The low correlations of the *HT* with the other tests indicates a substantial difference in the abilities being measured.

In Table 21, the test results are correlated with student variables. The statistical hypothesis, *that there is no correlation among the variables*, was tested with two-tailed test values of 0.088 and 0.155 for 5% and 1% respectively. Sixteen correlations were significant at the 1% level and a further four correlations were significant at the 5% level.

The *SCAT* correlated significantly at the 1% level with all except the *HT*. The grade level is significant, probably because the grade 7 students are on the fringe of the normal target population for this test. Age is significant for a similar reason. Sex is significant, probably because of the earlier maturation of girls at this stage of development. There also appeared to be some significance attached to the school attended — probably because not all schools had all three

TABLE 20

CORRELATIONS, MEANS AND STANDARD DEVIATIONS

FOR THE TEST BATTERY

TEST	*SCAT*	*IT*	*HT*	*GST*	*IT* and *HT*
SCAT	1.000	0.361**	0.044	0.656**	0.370**
IT		1.000	0.160**	0.306**	
HT			1.000	0.045	
GST				1.000	0.321**
IT and *HT*					1.000
Mean	59.96	27.61	64.18	18.60	91.80
Standard deviation	17.76	8.82	38.83	6.07	41.18

H_0: $r = 0$ $r_{obs}(0.05, 539) > 0.088$

$r_{obs}(0.01, 539) > 0.115$

**Significant at the 1% level.

TABLE 21

CORRELATIONS, MEANS AND STANDARD DEVIATIONS OF

TEST BATTERY FOR FOUR STUDENT VARIABLES

Variables/Tests	*SCAT*	*IT*	*HT*	*GST*	*IT* and *HT*
Grade	0.329**	0.244**	0.001	0.257**	0.247**
Age	0.194**	0.245**	0.080	0.241**	0.250**
Sex	0.757**	-0.108*	0.093*	-0.109*	0.197**
Means	58.96	27.62	64.19	18.60	91.81
Standard deviation	17.76	8.83	38.84	6.07	41.19

H_0: $r = 0$

*r_{obs} (0.05, 539) $\geq$ 0.088

**r_{obs} (0.01, 529) $\geq$ 0.115

grades and the presence or absence of grade 9 students made a difference in how well students performed on the test in the battery. In addition, the *SCAT* analysis shows a marked difference between schools so the difference in ability may show as a correlation. The school correlated at a low level but still significantly with the *HT*, indicating that there could be a difference in the teaching of the process skills. But since the teacher variables are outside the scope of this study, this question will have to be left unanswered. Sex showed a relatively high correlation with each test, particularly *SCAT*, but also correlating significantly (at the 5% level) with the other three tests. This high correlation with *SCAT* is probably linked to the difference in maturity level of girls as compared with boys at the same age level. Since the *SCAT* is relatively content-free this sex difference is more pronounced than on the other three tests. In the case of *SCAT*, 57.3% of the observed variance is tied to the sex variable. In the case of the other three tests only about 1% of the variance on each test is linked to the sex variable.

In Tables 20 and 21 the combined correlations of *IT* and *HT* were used to determine the relationship among age, sex, *SCAT*, *GST*, *IT* and *HT* simultaneously. The computations were made using:

$$R_{y \cdot x_1 x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2\, r_{yx_1}\, r_{yx_2}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}}$$

where $R_{y.xx}$ = coefficient of multiple correlation between y and a combination of x_1 and x_2

r_{yx_1} = the product-moment correlation between y and x_1

r_{xy_2} = the product-moment correlation between y and x_2

$r_{x_1x_2}$ = the product-moment correlation between x_1 and x_2

(from Popham and Sirotnik, 1967, p. 88).

As is indicated in Table 21 each of the multiple correlations are significant at the 1% level. This combination of the results of the *HT* and *IT* accounts for between 4% and 14% of the variance linked to each of the variables. Age is now linked to 6% of the variance of the two tests and is a significant factor in the analysis of the test battery. The school variable has a number of conditions linked to it, any one of which could influence the test scores: school size, class size, number of grades and type and quality of instruction. These, however, also are beyond the scope of the present study. These correlations strongly substantiate the concurrent validity of the *IT*, *HT* and *GST*.

A further test of the concurrent validity was made by conducting a stepwise regression analysis. In this stage of the analysis, the *GST*, *IT*, *HT* and *IT-HT* were each identified in turn as the criterion to determine the extent of the relationship of the various predictor variables.

In using and interpreting the results of the regression analysis one must be very conscious of the way in which predictor variables are chosen to enter the equation. That is, variables that correlate best are chosen first, but if two variables correlate well with the criterion and with each other, only one will be chosen. This means that a variable may be ignored, not because it is unimportant, but that another variable

accounts for as much of the variance as the ignored variable.

The relationships between tests are related to the internal relationships between and among items and this internal relationship can be expressed to some extent by its reliability coefficient. The K-R_{20} is a measure of internal consistency of a test and is somewhat vulnerable to pressures of time and to differences in quality among items. A further contribution to the interpretation of test reliability is in the relationship between the reliability coefficient and the standard error of measurement.

The standard error of measurement is an estimate of the standard deviation that would be obtained for a series of measurements of the same individual. This statistic taken together with the reliability coefficient probably provides the most satisfactory basis for comparing the tests used in the battery of instruments in this study. In comparing the scores achieved by an individual on all of the tests in this battery one must keep in mind the standard error of measurement of each test. The true score of an individual can be said to lie in a band two standard errors of measurement above and below the obtained score on a test. Two standard errors of measurement will contain 97.7% of the possibilities and represents quite a conservative estimate of the true score. In this case for the *SCAT* score an individual's true score likely lies in an area 15 points above or below his obtained score. In the case of the *GST* his true score is likely in a band 6.4

points above or below his obtained score. Using the reliability coefficient, the standard error of measurement can be computed using the formula:

$$S_m = S_t \sqrt{1 - r_{11}}$$

where S_m is the standard error of measurement

S_t is the standard deviation of the test scores

r_{11} is the reliability coefficient.

The reliability coefficients of these tests, presented in Table 22, are in an acceptable range and contribute to the overall validity of the tests. Reliability is a necessary condition for a test to have validity. It is the ceiling for the possible validity of the test. A test with a reliability coefficient of 0.00 is reflecting nothing but chance factors. It does not correlate with itself and cannot correlate with anything else. The theoretical ceiling for the correlation of the test with some other criterion measure is the square root of the reliability coefficient. Only to the extent that a test measures something accurately can it measure validly.

It is postulated that these tests have both the required reliability as shown by the reliability coefficients and the required validity as shown by the content validity and the descriptions of the factors being measured by the *IT*, *HT* and *GST*.

Stepwise Regression Analysis

A stepwise regression analysis was undertaken to determine the

TABLE 22

STANDARD ERRORS OF MEASUREMENT

Test	r_{11}	S_t	S_m
SCAT-V	0.763	11.86	5.76
SCAT-Q	0.647	8.10	4.81
SCAT-T	0.832	18.16	7.43
GST	0.820	7.47	3.18
HT	0.936	2.85	0.439
IT	0.845	0.845	0.119

interrelations that exist among the student variables and the test data. The important information from the analysis is included in Table 23. When the *GST* scores were identified as the criterion the *SCAT* total scores were the best predictors with the *HT* scores entering after the *SCAT*-V scores. Both sex and age are significant predictors of the *GST* scores. The sex weighting is negative indicating an inverse relationship. Age has the largest positive weight in the regression equation and is the single most significant factor in the equation.

In terms of the standard error of prediction, the *SCAT* score contributes the most with each of the variables reducing the error term by successively lower amounts. When converted to weights for the regression equation, though, the situation is reversed with the *SCAT* score having the lowest associated error term and sex having the highest. The *HT* score contributes only 1.5% to the predictability of the equation and 0.032 to the error term but reduces the overall error by a very modest 0.05.

When the combination of the *IT* and *HT* (*IT·HT*) scores are used as a criterion, the single most powerful predictor is the *SCAT*-V score. But this variable accounts for 87.8% of the total variance in the *IT·HT* scores, which were combined as a measure of propositional thinking and so identified in Table 24. As can be seen in Table 24, the accuracy of the prediction is also reasonably good with a total error term of 6.2 and a standard error of 0.023 associated with the regression weight. This strong relationship is taken as further evidence that the skills and abilities being measured by the *HT* and *IT* lie in the same realm of human skills and abilities being measured by the *SCAT*, more specifically those measured by *SCAT*-V.

TABLE 23

PREDICTORS OF *GENERAL SCIENCE TEST* SCORES

Predictor					Regression Weights			
Variables	F	P	% of Variance	Standard Error of prediction	Variables	Standard Weight	Weight	Standard Error
SCAT-T	405.85	0.000	43.04	4.586	Age	0.118	0.719	0.195
SCAT-T, Sex	24.87	0.000	45.57	4.487	Sex	-0.152	-1.846	0.378
SCAT-T, Sex, Age	12.29	0.000	46.79	4.441	*SCAT*-V	0.340	0.177	0.048
SCAT-T, Sex, Age, *SCAT*-V	9.06	0.002	47.41	4.408	*SCAT*-T	0.291	0.100	0.032
SCAT-T, Sex, Age, *SCAT*-V, *HT*	13.07	0.000	48.93	4.359	*HT*	0.120	0.188	0.032
						CONSTANT	-1.177	

Best Prediction Equation:

$$X_{GST} = -1.177 + 0.719\ X_{Age} - 1.846\ X_{Sex} + 0.177\ X_{SCAT\text{-}V} + 0.100\ X_{SCAT\text{-}T} + 0.188\ X_{HT}$$

TABLE 24

PREDICTORS OF *INFERENCE TEST* AND *HYPOTHESIS TEST* SCORES

Predictor					Regression Weights			
Variables	F	P	% of Variance	Standard Error of prediction	Variables	Standard Weight	Weight	Standard Error
SCAT-V	3869.67	0.0	87.814	6.206	*SCAT*-V	0.937	1.427	0.023
						CONSTANT	9.082	

Best Prediction Equation:

$$X_{IT \cdot HT} = 9.082 + 1.427\ X_{SCAT\text{-}V}$$

To further explore these relationships the last step of each item of the *HT* was scored independently as a subtest (*HA*). The balance of the *HT* was used as the criterion for a stepwise regression and the results are given in Table 25. The percentage of variance accounted for is so small as to be useless as a predictor but what is of interest in this analysis is the order and relative importance of the variables. The *HA* is the last part of each item in *HT*, the part which asks the student to predict a further event with the target. It is therefore not unrealistic that this is closely allied to the test as a whole. The next variables to enter are age and sex, indicating that they are related to the skills and abilities measured by the *HT*. The other variables are shown as they enter but are statistically insignificant beyond the 5% level. The regression equation is shown for interest but, as stated previously, has too high an error term to be a useful tool.

Taken in toto, the evidence from the regression analysis of the four tests indicates strong predictive relationships between *SCAT*-T and *GST*, *SCAT*-V and *GST*, *HT* and *GST* and *SCAT*-V and *IT*·*HT*. In the analysis used, there was no way to determine the influence of *IT* in the regression equations.

Summary of the Results of Validation

Two issues were raised during the course of this investigation. The first relates to the validity of the tests used in conducting the investigation and indirectly on the confidence that one may have in the data collected by those tests. The *content validity* of the *Inference* and the *Hypothesis Tests* rests totally upon the judgment of

TABLE 25

PREDICTORS OF *HYPOTHESIS TEST* SCORES

Predictor					Regression Weights			
Variables	F	P	% of Variance	Standard Error of prediction	Variables	Standard Weight	Weight	Standard Error
HA	45.770	.000	7.854	36.104	Age	-0.127	-4.821	1.599
HA, Age	10.969	.001	9.702	35.774	Sex	-0.087	-6.537	3.094
HA, Age, Sex	3.996	.046	10.371	35.674	*HA*	0.310	3.008	0.416
HA, Age, Sex, *GST*	1.497	.22	10.622	35.658	*GST*	0.054	-0.333	0.272
HA, Age, Sex, *GST*, *IT*					CONSTANT	131.489		
HA, Age, Sex, *GST*, *IT*, Grade								
HA, Age, Sex, *GST*, *IT*, Grade, *SCAT*								

Best Prediction Equation:

$$X_{HT} = 131.489 - 4.821\ X_{Age} - 6.537\ X_{Sex} + 3.008\ X_{HA} - .333\ X_{GST}$$

the panel as to how well the tests measure the described behaviors that serve to define the tests. It has been reported that these judgments are favorable.

The *construct validity* depends upon the factor structure identified as underlying the tests and their definition in terms of the behaviors that define the abilities they are supposed to measure. The evidence strongly supports the contention for construct validity.

The *concurrent validity* of the *Inference Test* and the *Hypothesis Test* depends upon correlation with two additional instruments. The mental process dimension of the two tests correlates, within limits, with the *Cooperative School and College Ability Test* which would indicate that the cognitive functions called upon by the two tests, albeit to differing extents, are related to the scholastic abilities being measured by *SCAT* and more particularly *SCAT*-V, the verbal subtest. The scientific process dimension of the two tests correlates with the *General Science Test* developed specifically to measure the student's knowledge of and use of the inferring and hypothesizing skills. The establishment of the validity of this test was treated as a somewhat separate question earlier. The concurrent validity of the *Inference Test* and *Hypothesis Test* was further supported by the use of the regression equation.

The validity of the *Inference Test* and *Hypothesis Test* has significant support from these four sources. In addition, the test

statistics tend to show a relatively stable and coherent structure with the value of the Kuder-Richardson coefficients which show that there is evidence to address the issue in favor of the validity of the two tests.

The second issue relates to the hypotheses stated in Chapter I and this will now be addressed.

Test of Stated Hypotheses

Each of the hypotheses posed in Chapter I will be considered in turn.

Hypothesis H_1: *There is no significant difference between the mean score on the* IT *among boys and girls in age categories from 11 to 15.*

To test this hypothesis, the data presented in Table 26 were subjected to a check of the homogeneity of variance, an analysis of variance, and a multiple comparison of group means using the Scheffé procedure. This method uses the criterion that the probability of rejecting the null hypothesis when it is true should not exceed 0.01 or 0.05 for any of the comparisons. The procedure involves the calculation of the F ratio for each comparison, determination of the F of the desired confidence level ($F_{0.99}$) for d.f. = (number of comparisons (k) -1) and (number of subjects (N), number of comparisons (k) and calculation of F') where $F' = (k-1)F_{0.01}$. For any difference to be significant at the required level, F must be greater than or equal to F'. The ANOV 15 program computed the F and F' and returned a table of the level of significance for each comparison among the

TABLE 26

STUDENT PERFORMANCE ON *INFERENCE TEST*

Age Group			N	MEAN	VARIANCE
1	Boys	12 and under	38	25.21	86.28
2		13	71	27.03	70.31
3		14	99	29.72[1]	40.45
4		15 and over	48	31.50[2]	30.16
5	Girls	12 and under	44	23.05[1 2 3]	116.84
6		13	84	25.08	96.41
7		14	100	27.83	90.37
8		15 and over	55	30.31[3]	58.81
		TOTAL	539	27.46	77.78

NOTE: Superscripts indicate that significant differences exist between those groups with matching numbers.

group means. The data were tested for *homogeneity of variance* using the χ^2 statistic. The value returned for the data was $\chi^2 = 42.00$ with a 0.0 probability that it was by chance. The variance is different from group to group. Thus assured of difference one can proceed with the next step of testing the *analysis of variance* by finding the ratio of the mean square of the groups and that due to error. This resulted in an *F* ratio of 4.87 which is significant beyond the 1% level. That is to say the group means are not equal. The ANOV 15 program then returned the probability matrix for a Scheffé multiple comparison of means. The differences between the means of group 3 and 5 (6.67), 4 and 5 (8.45) and 5 and 8 (7.26) were identified as being significant beyond the 5% level. That is, there is a significant difference between the performance of the group of young girls (group 5) and the older boys and the eldest girls. That is not to say there are no differences among the other groups' mean scores but that the differences are not statistically significant at the 5% level of confidence. In terms of H_1, the hypothesis must be rejected as stated.

The significant differences in the *IT* scores resulted from a difference in the performance of the age 12-and-under girls when compared with 15-and-older girls and the 14-and-older boys. Other differences are apparent in the data and are in the expected direction but are not statistically significant. One reason for the low number of significant differences is probably related to the Scheffé test which is very conservative but powerful in that there is a smaller chance of making Type I errors (rejecting a null hypothesis in error). It is clear from the data presented in Table 26 that there is some

increase in competence in making inferences with increased maturity.

Hypothesis H_2: *There is no significant correlation between the student scores on the* IT *and the school attended, age, sex, grade,* SCAT *score,* HT *score or* GST *score.*

To test this hypothesis reference was made to Table 21 where it was found that correlation of *IT* with the school attended is 0.195 — low but still significant at the 5% level of confidence. The low correlation is probably related to variations in the general level of scholastic ability, quality of teaching and other uncontrolled classroom conditions. The correlations of *IT* with age and grade are 0.245 and 0.244, respectively, and are larger than that with the sex variable (-0.108). All of the correlations with student variables are statistically significant. Reference was then made to Table 20 where the correlations of *IT* with *SCAT*, *HT* and *GST* were found to be 0.361, 0.160 and 0.306, respectively. These are significant at the 1% level.

The low correlation with *HT* was somewhat disappointing but probably reflects the increased complexity of the hypothesizing skill. The correlation of *IT* with *GST* was at a more satisfactory level (0.306). This reflects the similarities in the cognitive level being measured. Hypothesis H_2 is rejected as stated.

The correlation of *IT* with student variables and other tests in the battery are all significant at the 1% level of confidence — with the exception of sex which is at the 5% level. It was expected that there would be significant correlations between *IT* and the other tests since it was suggested that *SCAT* and *IT* both test related cognitive

levels and that *GST* and *IT* are measuring related scientific processes. The low correlation between *IT* and *HT* is due to the degree of difference that exists between *IT* and *HT*. The fact that there *is* any significant correlation at all is probably because of the hierarchical relationship that exists between the two skills.

Hypothesis H_3: *There is no significant difference between the mean score on the* HT *among boys and girls in age categories from 11 to 15.*

To test this hypothesis the data in Table 27 were subjected to a check of the homogeneity of variance, an analysis of variance and a multiple comparison of the group means.

The data were tested for *homogeneity of variance* using the χ^2 statistic. The computer program returned a χ^2 for these data of 8.153 which is less than the critical value ($\chi^2_{0.01}$ (7) = 18.5) for significance at the 1% level. That is, the variances are not significantly different.

The *analysis of variance* resulted in an *F* test of the ratio of the mean square of the group and that due to error. The resulting *F* ratio was 1.51, which is less than the critical value needed for significance at the 1% level. That is, there are no significant differences among the means, although there are observable differences which are in the expected direction; that is, the means increase with age category. H_3 is not rejected and remains as stated.

There were no significant differences in the mean scores of the age groups on *IT*. The reason that the observed differences are not significant is probably related to the large dispersion of the scores

TABLE 27

STUDENT PERFORMANCE ON *HYPOTHESIS TEST*

Age Group			N	MEAN	VARIANCE
1	Boys	12 and under	38	56.38	1302.4
2		13	71	58.63	1318.8
3		14	99	63.11	1431.3
4		15 and over	48	67.39	1781.2
5	Girls	12 and under	44	54.92	1661.8
6		13	84	57.96	1332.5
7		14	100	59.73	1234.6
8		15 and over	55	67.45	1505.8
		TOTAL	539	60.27	1446.7

which resulted in a high variance among the test scores. This variance is probably related to the difficulty level of the test (.357). The observable differences in the data are in the expected direction which is that competence in hypotheses formulation increases with age.

Hypothesis H_4: *There is no significant correlation between student scores on the* HT *and the age, sex, grade,* SCAT *score or* GST *score.*

To test this hypothesis reference was made to Tables 20 and 21 where it was found that the correlation of *HT* with age category (0.080) and grade level (0.001) are not significant. The correlation with sex (0.093), which indicates little relationship, is still significant at the 5% level. The correlation of *HT* with *SCAT* is low (0.044) as is that with *GST* (0.015). The only significant correlation of *HT* is with sex, so hypothesis H_4 must be rejected for the sex variable. The rest of the hypothesis is not rejected as stated.

The correlation of *HT* with student variables and other tests in the battery are very low and only correlate significantly with *IT* (see hypothesis H_2) and the sex variable. The low correlation with *IT* has been discussed under "concurrent validity" and hypothesis H_2 above. The sex correlate may well indicate an increased cognitive maturity of girls at this age level.

Hypothesis H_5: *There are no significant differences between the mean scores of boys and girls on the combined* IT *and* HT *as an indicator of propositional thinking, and their age category from 11 to 15.*

To test this hypothesis the data in Table 28 were subjected to two analyses, first a test for homogeneity of variance and second, the analysis of the variance. The data were tested for *homogeneity of variance* using the χ^2 statistic, the value returned by the computer program being $\chi^2 = 4.34$, which was less than the critical value needed for significance. Therefore there is no significant difference in the variance.

The *analysis of variance* was tested by comparing the mean square of the groups with that due to error. The resulting F ratio was 1.20 which is less than the critical value needed for significance. There are no grounds on which to reject H_5 although, by inspection, the differences are in the expected direction.

The insignificance of the mean differences of the groups on the combined *IT·HT* is probably related to the same problem with the large dispersion of *HT* scores already discussed. The observable differences are in the expected direction and, given a more satisfactory student performance on the *HT*, more of the mean differences might gain significance.

The lack of observable increase in competence in the combination of inferring and hypothesizing skills with age is contrary to the findings of Bruner and Gagné as cited in Chapter II of this report.

TABLE 28

STUDENT PERFORMANCE ON THE COMBINATION OF

INFERENCE TEST and *HYPOTHESIS TEST*

Age Group			N	MEAN	VARIANCE
1	Boys	12 and under	38	92.57	1710.5
2		13	71	93.83	1583.6
3		14	99	96.26	1681.6
4		15 and over	48	101.32	2059.3
5	Girls	12 and under	44	85.08	1953.7
6		13	84	86.61	1664.8
7		14	100	89.38	1517.0
8		15 and over	55	92.89	1813.0
		TOTAL	539	91.81	1693.3

Hypothesis H_6: *There is no significant correlation between student scores on the combined* IT *and* HT, *as an indicator of propositional thinking, and age, sex, grade,* SCAT *score and* GST *score.*

To test this hypothesis reference was made to Tables 20 and 21 where it was found that the *IT·HT* correlates at the 1% level of significance with age, sex, grade, *SCAT* score, *GST* score with values of 0.250, 0.197, 0.247, 0.370 and 0.321, respectively. Since all of the correlations are significant at the 1% level of confidence, H_6 is rejected as stated.

The correlation of *IT·HT* with student variables and other tests in the battery are all significant at the 1% level of confidence. The correlations are all quite low, but indicate that the combined score explains more of the variance than either *IT* or *HT* independently. The combined test seems to provide, as indicated by the improved correlations, a more complete picture of the ability of students to infer and hypothesize than either of the individual tests. This conclusion is compatible with the basic premise of this study, that the cognitive maturity results in an increased competence in the ability to make inferences and to formulate hypotheses.

Hypothesis H_7: *There is no significant difference between the mean scores of boys and girls on the SCAT, and the school attended and age.*

To test the first part of H_7 reference is made to Table 19. A comparison was made among the mean scores of each of the 11 schools and four significant differences were discovered. Schools 2, 9 and 11 were different at a 1% level of significance and school 6 at a 5% level. To test the second part of H_7, data from Table 29 is subjected to three separate analyses: first of all a check on the homogeneity of variance, then an analysis of variance and finally a comparison of the group means.

The data were tested for *homogeneity of variance* using the χ^2 statistic, the value returned for this data being $\chi^2 = 3.62$ which is less than the critical value needed for significance. Therefore, the variances are not significantly different.

The *analysis of variance* was tested by determining the ratio of the mean square of the groups and that due to error. This resulted in an F ratio of 5.16 which is greater than the critical value needed for significance. There is a difference among the group means.

The *comparison of means* was carried out using the Scheffé procedure and it was found that one pair of means differed significantly, those of groups 1 and 7. The difference between the youngest boys and older girls amounted to a difference of 19.25 which is significant at the 5% level. The next smallest difference between the youngest boys and girls (19.17) was only significant at the 12% level. On this basis H_7 must be rejected in both parts.

TABLE 29

STUDENT PERFORMANCE ON *COOPERATIVE SCHOOL AND COLLEGE ABILITY TEST*

Age Group			N	MEAN	VARIANCE
1	Boys	12 and under	38	51.13	288.01
2		13	71	55.42	310.85
3		14	99	62.15	320.93
4		15 and over	48	57.79	294.32
5	Girls	12 and under	44	52.91	269.95
6		13	84	55.31	271.16
7		14	100	65.08	284.93
8		15 and over	55	64.83	278.15
		TOTAL	539	58.96	314.9

The differences among the mean scores on *SCAT* are related only to the youngest boys when compared with the older girls. The only unexpected data from the *SCAT* test were the better than average performance of the 14-year-olds in the sample. Both boys and girls in this group did substantially better than expected. This information, coupled with the significant differences from four schools in the sample, suggests that these schools had an atypical group of 14-year-old students in the year in which the study was completed. However, this deviation from the homogeneity of the groups did not appear to have a great effect on the analysis of variance tests. The *F* test is relatively robust with respect to deviations from the basic assumptions with respect to the homogeneity of the sample.

Hypothesis H_8: *There is no significant difference between the mean scores of boys and girls in age categories of 11 to 15 on the* GST.

To test this hypothesis, data from Table 30 were subjected to a check of the homogeneity of variance, an analysis of variance, and a multiple comparison of the group means.

The data were tested for *homogeneity of variance* using the χ^2 statistic, the value returned for this data being $\chi^2 = 28.77$ which is greater than the critical value needed for significance. The variances are significantly different.

The *analysis of variance* was tested by determining the ratio of the mean square of the groups and that which was due to the error. This resulted in an *F* ratio of 6.64 which is greater than the critical value

TABLE 30

STUDENT PERFORMANCE ON *GENERAL SCIENCE TEST*

Age Group			N	MEAN	VARIANCE
1	Boys	12 and under	38	16.47[1]	21.18
2		13	71	17.39[2]	30.21
3		14	99	21.55[1234]	49.13
4		15 and over	48	19.95	53.70
5	Girls	12 and under	44	15.34[3]	19.25
6		13	84	16.86	23.51
7		14	100	19.38[4]	30.56
8		15 and over	55	19.23	34.14
		TOTAL	539	18.60	36.79

NOTE: *Subscripts indicate that significant differences exist between those groups with matching numbers.

needed for significance, that is, there is a difference among the group means.

The multiple *comparison of means* was carried out using the Scheffé procedure. It was found that four pairs of means differed significantly at the 1% level: groups 1 and 3, groups 2 and 3, groups 3 and 5, and groups 3 and 7. On this basis H_8 is rejected as stated.

The differences among the mean scores on *GST* indicate that the group of 14-year-old boys out-performed every other group taking the test. Otherwise the mean differences are much as one would expect. That is, the competence in, and familiarity with, scientific processes improves somewhat with age but is probably more closely related to the teaching of science.

Summary of the Testing of the Hypotheses

H_1 is rejected in toto.

H_2 is rejected in toto.

H_3 is not rejected.

H_4 is rejected for school and sex variables and by implication from H_2 for *IT* score, but not for age, grade, *SCAT* score, or *GST* score.

H_5 is not rejected.

H_6 is rejected in toto.

H_7 is rejected in toto.

H_8 is rejected for four mean differences only.

The main purpose of the study was to provide some insights into the inferring and hypothesizing abilities that students exhibit in the

classroom, that is, the extent to which students can think propositionally. Evidence has been presented to the effect that most students in the sample can make inferences.

Students also appear to grow in their capacity to formulate hypotheses as their intellect matures. The evidence for this growth in capacity is not unequivocal but is marred by the wide range in the scores and wide variance of the groups sampled. There are a number of reasons that could account for this disparity — differences in the classroom conditions, variations in the testing conditions, variations in the quality of instruction that students have had, and variations in their attitudes toward the teacher and the tests. The combined *IT* and *HT* tests form a more powerful tool than either test individually, inasmuch as the correlation with student variables was improved and there was an increase in amount of the variance accounted for in the *SCAT* and *GST* scores. In terms of the question posed as part of the problem in Chapter I of this study, the ability to think propositionally is not firmly established in the skill repertoire of junior high students. By inspection of the data, the sharpest break in the difference between mean scores on the *HT* appears to be between the 14- and 15-year-old girls, with no such break in the boys' mean scores. The search for broad trends would only be fruitful with a much larger sample of students.

The second question posed, with respect to the relation of scholastic ability and knowledge of scientific processes, is reasonably clear. There is a definite correlation between inferring ability and students' general scholastic ability as measured by *SCAT* and their

knowledge of and ability to use selected scientific processes as measured by *GST*. The ability to hypothesize is not correlated with *SCAT* and *GST* in any significant sense. The reason for this is probably related to: 1) students have not developed the formal operational cognitive skills, 2) students are not taught to use their skills to formulate hypotheses in any but contrived problems, and under controlled situations, and 3) students who have recently acquired a skill often are unable to practice it in a novel or different situation and revert to a previous level of function when faced with a stressful situation.

For these or other reasons, students in the study sample did not exhibit any marked skill in formulating hypotheses.

When the results of the combined *IT·HT* are examined it is apparent that there is a steady improvement in students' ability to think propositionally but there is no sudden shift in the mean scores which would indicate a greatly improved capacity to deal with problems on a higher intellectual plane. One could only suggest that the trends in the data indicate that further study involving a wider age range and broader geographic base is needed to collect more data pertaining to students' capacity to make inferences and to formulate hypotheses.

CHAPTER V

SUMMARY, CONCLUSIONS, LIMITATIONS, IMPLICATIONS FOR SCIENCE EDUCATION, AND IMPLICATIONS FOR FURTHER RESEARCH

Purpose of the Investigation

The purpose of the investigation was to make a contribution to the developing theory of science learning and to provide some insight into the skills and abilities that students use in the science classroom. Specifically, the study was designed to provide information pertaining to the following questions:

1. *How is the ability of propositional thinking as defined in terms of the ability to use the scientific processes of inferring and hypothesizing distributed among the student population (by age, by grade and by sex)?*
2. *Is the ability to think propositionally related to a student's scholastic ability and knowledge of scientific processes?*

The present study defined the acquisition of the ability to formulate hypotheses as being a manifestation of the onset of formal operations, and equated this with the Piagetian propositional thinking. In addition, the formation of hypotheses was defined as being hierarchically related to the concrete operations of observing and inferring in a sequence of: Observing → Inferring → Hypothesizing.

The Findings from the Review of the Literature

The review of the literature confirms the solid place that the process dimension has attained in the teaching of science. The many attempts at defining the dimensions of scientific processes may result in minor differences in specific definitions but there is substantive agreement on the importance of the ability to abstract the essential parts of observed phenomena and formulate hypotheses. In two words: *think propositionally*. There is also a substantial amount of agreement as to the developmental and hierarchical relationships involved in the acquisition and refinement of this ability.

Piaget (1964) describes this ability to think in a propositional fashion as part of a total view of learning employing such constructs as schemata, assimilation, accommodation, operations, reversibility and equilibration. Specifically, the ability to think propositionally is closely linked with Piaget's stage of formal operations which is characteristically attained by the ages of 10 to 12.

The Brunerian concept of recurring learning cycles which carry the learner to higher levels of abstraction and generalization can be viewed as a useful extension of Piaget's theory of cognitive growth. Bruner's (1973) description of concept development is useful in the development of descriptions of children's classroom behavior.

The review of existing tests and approaches to the measurement of scientific process related abilities led the researcher to the development of a pencil and paper test that is an extension of a

game-type activity developed by Michiche and Keany (1969). This in turn led to a concern with the test-development tasks and the necessary emphasis on validation and test statistics.

Summary of the Test Results

Students from 30 classrooms participated in the main part of the investigation. In May and June, 1972, a battery of four tests was administered to the students. The battery consisted of the *Cooperative School and College Abilities Test (SCAT)*, *General Science Test (GST)*, *Inference Test (IT)*, and *Hypothesis Test (HT)*.

The results of the *IT* and *HT* were processed by DERS programs DEST 02, NONP 10 and FACT 01. The information for the *IT* shows a satisfactory reliability (0.845) and a highly unidimensional substructure with a slight relationship to age and scholastic ability (4.0% and 2.6% of the total variance). The other test statistics showed a satisfactory level of item difficulty, mean and standard deviation.

For the *HT* the reliability coefficient of 0.936 may be optimistic because of the difficulty level of the items. The factor analysis also showed a relatively simple structure with some indication of a relationship with age category and grade level, and sex and scholastic ability (6.3% and 4.3% of the total variance in each case).

The results of the *GST* were processed by DERS programs TEST 01, DEST 02, and FACT 01. The test statistics are satisfactory. The reliability (0.820) indicates either a relatively simple factor structure or a relatively close correlation among items; in other words,

a consistency within the test as to the things being measured. The mean and variance (21.0 and 55.84, respectively) indicate a difficult test which is to be expected since the items chosen were of the synthesis and evaluation levels of Bloom's taxonomy. The factor analysis showed a surprisingly complex structure and repeated attempts to derive a simpler structure failed to diminish the number of factors. The 19 factors all relate to the processes of science and vary from each other only in level or in focus.

The *SCAT* 3B was processed by the Department of Education Student Evaluation and Data Processing Services Branch scoring program. The results have been reported previously and show a satisfactorily high reliability of 0.832. The variance proved to be homogenous at the 1% level but there proved to be significant differences in scholastic ability from school to school. In this regard, then, any significance that was attached to the school attended by students has been interpreted as being differences in scholastic ability. The low but statistically significant correlation lends some credence to this practice.

The *content validity* of *IT* and *HT* was judged by a panel of nine teachers and nine students on the grounds of 1) clarity of instructions, 2) difficulty of the items, 3) interest in the format, and 4) relationship of items to a behavioral definition. The judgment was favorable.

The *construct validity* was measured in terms of the simplicity of the factor structure and the relation of that structure to the basic

definitions established for inferring and hypothesizing. These factor structures strongly supported the contention of construct validity.

The *concurrent validity* of *IT* and *HT* was established by determining the correlation of these two tests with *SCAT* to establish a link with scholastic ability, but more particularly to establish a link with the selected scientific processes. The *IT* has a good base for claiming concurrent validity while the *HT* has a less strong basis.

On balance, the claim of validity for and hence the confidence that one can place in the data gathered by the tests has been substantiated. One can have confidence that the hypotheses that were posed and tested have been done so using valid data from a widely disparate but substantial sample of junior high school students.

As with much educational research *in situ*, the results are not as clear and unequivocal as one would wish but there are some indicators in this study.

The excellent response of students to the *IT* and the novelty of the puzzle-format of the *IT* would recommend it as an instrument for evaluating student skills in making inferences in isolation from the factual base of a real-life situation — which is often desirable. It is likely from the test statistics that the test could be used without modification at the grades 5 and 6 levels as well.

The test statistics from the *HT* would indicate that the test could be used at a higher grade with little modification. The students found the test quite difficult — because of the test format or because

of the skills being tested is difficult to isolate. In combination with the *IT* though, the two tests become good measures of the same scholastic abilities that are involved in the *SCAT* verbal test.

Summary of the Hypotheses Tests

The linkage of the scholastic abilities and propositional thinking as it is being used in this study lend further support to the teaching of the skills of formulating hypotheses in the science classroom as a means of helping students acquire the broad learning skills necessary to succeed as adults. These skills are necessary for flexible, creative approaches to the solution of novel problems. The concept of propositional thinking as it is being used here seems to arise naturally from Piagetian developmental theory and to fit into Bruner's cyclic learning sequence and Gagné's process hierarchy very well.

On the basis of the stated hypotheses and the statistical analysis of the data to test them the questions posed at the beginning of this report can be answered as follows:

1. The ability to think propositionally is closely related to the skill in using the science process skills as measured by the *General Science Test*.

2. The ability to think propositionally is closely related to the general scholastic abilities measured by the *SCAT* verbal test as evidenced by the correlations between the

two tests and the predictability of the combined *HT* and *IT* scores by the *SCAT* verbal scores.

3. The student's age category is a minor but significant contributor to the explanation of the variance in the combined *IT* and *HT* and is therefore a significant factor in the ability to think propositionally.

4. The sex of a student is also a factor in the ability to think propositionally but is even more important in the area of scholastic ability in junior high school.

It is apparent that the maturation difference between adolescent boys and girls has had a significant effect on their ability to think propositionally. Whether this is a true difference in ability levels in favor of the boys or a temporary and ephemeral difference among individuals that can be explained in terms of a transitional state in the evolution of the formal operation stage, or a cultural difference, is an open question.

In general, one can reasonably maintain that the role of propositional thinking in using scientific processes has been supported by the multiple correlation of the *Hypothesis Test* and *Inference Test* with the *General Science Test*. In terms of the evidence it would follow that the concept of propositional thinking is a useful construct in the teaching of scientific processes.

It would also appear that there is much to be gained by teaching the skills of making inferences and formulating hypotheses in the scientific fields because the generalized abilities so formed have a much broader scope of scholastic abilities. Such teaching should facilitate the individual's progression to higher levels of abstraction and to a stronger grasp of cognitive structures that are evolving concurrently. They should also lead to a more generalized ability to deal with new (to the student) thought patterns and to achieve equilibration sooner and with less confusion.

There was no clear distinction among groups of individuals to indicate whether a specific group had clearly achieved equilibrium at the formal operations stage. This general inability for many students to formulate hypotheses corroborates other studies in this area (Hobbs, 1972, p. 126) that have concentrated on other manifestations of formal operations. This corroboration extends to the "edge" that boys appear to have in attaining this level of cognition, and the general "raggedness" with which students achieve this level.

Implications for Science Educators

In the interests of clarity and brevity, implications for science educators arising from the present study are offered in point form.

1. In view of the importance of propositional thinking in the use of scientific processes, every student should be taught as soon as it is practicable to make inferences from observations made of real life situations in addition to artificial, controlled laboratory situations.
2. Students should be taught the skills of formulating hypotheses from a number of inferences as soon as they exhibit the capacity of generalizing from concrete situations.
3. Scientific processes should be taught in a sequence of basic concepts and operations built in an unbroken hierarchy consistent with the developing intellect of the student so that the more complex skills are being developed in step with the child's ability to generalize, symbolize and formalize. The child should develop a genuine intuitive feeling for scientific process skills so as to take full advantage of the generalized scholastic abilities that are involved with propositional thinking.
4. The importance of realizing what modes of thought and levels of abstraction and generalization are available to one's students cannot be overemphasized.
5. Different modes of teaching scientific process skills should be investigated — it seems that many science teachers act as though students learn these skills by reading and talking about them. That is, as much care should be devoted to the teaching of process skills as to the content of science courses.
6. Students should be encouraged to use their inquiry skills learned in the science class in a variety of situations, that is, making

inferences and formulating hypotheses about other than science-related activities.

7. In structuring a total science program from K through 12, recognition should be given to the writings of Piaget, Bruner and Gagné inasmuch as one should recognize that the developing intellect of the individual can be encouraged, strengthened and guided by a structural sequence of learning based in the structure of the scientific process hierarchy.

Implications for Further Research

One of the most pressing needs in the further development of theory of science education is that for a longitudinal study of the intellectual abilities of children as they are exhibited in the science program. Such a study, by following a group of children through the years, would shed additional light on how children's intellectual development is aided or impeded by the requirements of the science program.

Another obvious extension of the present study is the study of the integrated processes as they are used by senior high school students. The increasing pressures on high school students, both from the curriculum and the community, demand that educators become more sensitive to both the time commitments and the prerequisite skills being demanded.

It would be instructive and informative to investigate the practice effects involved in the rewriting of the *HT* and *IT* to see if the tests are stable over a number of administrations. It would also be

interesting to investigate the effects on a student's inferring and hypothesizing skills that might accrue from writing the tests. *Does the type of reasoning that is required in writing the tests give the student insights into how to infer or to formulate hypotheses?*

BIBLIOGRAPHY

AAAS, Commission on Science Education. 1965. *The Psychological Bases of Science — A Process Approach.* Washington: AAAS, Misc. Publ. 65-8.

__________. 1965. *An Evaluation Model and its Application.* Washington: AAAS, Misc. Publ. 65-9.

__________. 1967. *Science — A Process Approach, Purposes, Accomplishments, Expectations.* Washington: AAAS, Misc. Publ. 67-12.

__________. 1968. *Commentary for Teachers.* Washington: AAAS, Misc. Publ. 68-7.

Alberta Department of Education. 1969. *Curriculum Guide for Elementary School Science.* Edmonton: Queen's Printer.

__________. 1974. *Curriculum Guide for Junior High School Science.* Edmonton: Queen's Printer.

Alpren, M. 1946. The ability to test hypotheses. *Science Education* 30: 220-229.

Anderson, R. 1965. Children's ability to formulate mental models to explain natural phenomena. *Journal of Research in Science Teaching* 3: 326-332.

Atkin, M.J. An analysis of the development of elementary school children in certain selected aspects of problem solving ability: A study of formulating and suggesting tests for hypotheses in science learning experiences. Unpublished Ph.D. Dissertation, New York University, 1956.

__________. 1958. A study of formulating and suggesting tests for hypotheses in elementary school science learning. *Science Education*

42: 414-422.

____________, and Karplus, R. 1962. Discovery or invention. *Science Teacher* 29(5): 45-51.

Ausubel, D. 1964. The transition from concrete to abstract functioning. *Journal of Research in Science Teaching* 2: 261-266.

____________. 1967. *Learning Theory and Classroom Practice*. Toronto: Ontario Institute for Studies in Education.

____________. 1968. *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart and Winston.

Ayers, J.D. 1967. *Test Item Construction: A Manual for Teachers*. Edmonton: Alberta Teachers' Association.

Barker, R. Development and evaluation of a method of structuring inquiry films to elicit hypotheses from high school chemistry students. Paper presented to the 43rd Annual Meeting of the National Association for Research in Science Teaching, Minneapolis, Minn., March 6, 1970.

Belanger, M. 1969. Learning studies in science education. *Review of Educational Research* 39(4): 377-395.

Beyer, B. 1968. The mystery of inquiry teaching. *Ohio Schools* 46: 18-20.

Beard, J. 1971. Development of group achievement tests for two basic processes of AAAS — A process approach. *Journal of Research in Science Teaching* 8: 179-183.

Blackford, D. The classification abilities of elementary school students. Unpublished Master's Thesis, University of Alberta, 1970.

Bloom, B.S., Hastings, J.T., Madaus, G.F. 1971. *Handbook on Formative and Summative Evaluation of Student Learning.* New York: McGraw-Hill.

Brakken, W.W. 1967. Meaning of process approach. *Instructor* 76: 21.

Brown, W.R. 1968. Defining the processes of science: A statement by educators in India. *The Science Teacher* 35(9): 26.

Bruner, J.S. 1961. The act of discovery. *Harvard Educational Review* 31(1): 21-32.

___________. 1973. *Beyond the Information Given.* Edited by Jeremy M. Anglin. New York: Norton.

___________, Goodnow, J.J. and Austin, G.A. 1956. *A Study of Thinking.* New York: Wiley and Sons.

Buros, O.K. 1959. *The Fifth Mental Measurements Yearbook.* Highland Park: Gryphon Press.

___________. 1965. *The Sixth Mental Measurements Yearbook.* Highland Park: Gryphon Press.

Butts, D.P. 1963a. The degree to which children conceptualize from science experience. *Journal of Research in Science Teaching* 1: 135-143.

___________. 1963b. Does experience equal understanding? *Science Teacher* 30(8): 81-82.

___________. 1964. The evaluation of problem solving in science. *Journal of Research in Science Teaching* 2: 116-122.

___________. 1965. The relationship of problem solving ability and science knowledge. *Science Education* 49: 138-146.

__________, and Jones, H. 1966. Inquiry training and problem solving in elementary school children. *Journal of Research in Science Teaching* 4: 21-27.

Byers, J. 1965. Hypothesis behavior in concept attainment. *Journal of Educational Psychology* 56: 337-342.

Coleman, C. 1966. A demonstration of alternative hypotheses. *School Science and Mathematics* 66: 81-85.

Coleman, J.S. 1961. *Adolescent Society*. Glencoe, Ill.: Free Press.

Connelly, F. 1970. The role of principles of enquiry in the conduct of enquiry and in curriculum development. *Science Teacher* 37(9): 23-26.

Copeland, R.W. 1970. *How Children Learn Mathematics*. Toronto: Collier-Macmillan.

Downey, L.W. 1965. *The Secondary Phase of Education*. New York: Blaisdell.

Educational Policies Commission. 1966. *Education and the Spirits of Science*. Washington, D.C.: National Educational Association.

Elkind, D. 1961. Quantity conceptions in junior and senior high school students. *Child Development* 32: 251-260.

__________. 1970. *Children and Adolescents: Interpretive essays on Jean Piaget*. New York: Oxford University Press.

__________, and Flavell, J. 1969. *Studies in Cognitive Development*. New York: Oxford University Press.

Farre, G.L. 1966. On the problem of scientific discovery. *Science Teacher* 33(7): 26-29.

Ferguson, G.A. 1966. *Statistical Analysis in Psychology and Education.* New York: McGraw-Hill.

Flavell, J.H. 1963. *The Developmental Psychology of Jean Piaget.* New York: Van Nostrand.

Frederiksen, N. 1959. Development of a formulating hypothesis test. In: *The 1958-59 Annual Report of the Educational Testing Service.* Princeton: Educational Testing Service.

Gagné, R.M. 1963. The learning requirements for enquiry. *Journal of Research in Science Teaching* 1: 144-153.

___________. 1968. Contributions of learning to human development. *Psychological Review* 75(3): 177-191.

___________. 1974. Educational technology and the learning process. *Educational Researcher* 3(1): 3-8.

Geisinger, R.W. Inquiry training: Results in tenth grade students of complex heuristic treatments. An unpublished Doctoral Dissertation, University of Pittsburgh, 1964.

Gibbs, R. An analysis of the effectiveness of biological science curriculum student single topic films in teaching hypothesis construction to high school biology students. Unpublished Ph.D. Dissertation, Indiana University, 1967.

Guilford, J.P. 1956. The structure of intellect. *Psychological Bulletin* 53: 267-293.

Hanson, N. 1967. An anatomy of discovery. *Journal of Philosophy* 64: 321-352.

Hedges, W. 1966. *Testing and Evaluation for the Sciences.* Belmont: Wadsworth Publishing.

Hobbs, E.E. Conservation of weight and volume in secondary school

students. Unpublished Master's Thesis. University of Alberta, 1972.

Hurd, P. DeH. 1964. *Toward a Theory of Science Education Consistent with Modern Science, Theory into action . . . in Science Curriculum Development.* Washington: National Science Teachers Association, pp. 5-15.

__________. 1969. *New Directions in Teaching Secondary School Science.* Chicago: Rand McNally.

Inhelder, B. and Piaget, J. 1958. *Growth of Logical Thinking from Childhood to Adolescence.* New York: Basic Books.

Jackson, S. 1965. The growth of logical thinking in normal and subnormal children. *British Journal of Educational Psychology* 35: 255 cf.

Karplus, R., and Thier, H.D. 1969. *A New Look at Elementary School Science.* Chicago: Rand McNally.

Klein, S.P. 1969. Anxiety and learning to formulate hypotheses. *Journal of Educational Psychology* 60: 465-475.

Kuypers, D. The development and utilization of hypotheses by children. Unpublished Ph.D. Dissertation, University of Illinois, 1969.

La Bohn, W. 1966. So you want to provide a modern science program. *School Science and Mathematics* 65: 698.

Lavatelli, C.S. 1970. *Piaget's Theory Applied to an Early Childhood Curriculum.* Boston: American Science and Engineering.

__________. 1971. Piaget's developmental theory of learning and its implications for instruction in science. In: *Readings in Science Education for the Elementary School*, 2nd ed. Edited by E. Victor and M.S. Lerner. New York: Macmillan.

Livermore, A. 1964. The process approach of the AAAS Commission on Science Education. *Journal of Research in Science Teaching* 2(4): 271-282.

Lovell, K. 1961. A follow-up study of Inhelder and Piaget's *The Growth of Logical Thinking*. *British Journal of Psychology* 52: 143-153.

Maguire, T.O. and Hazlett, C.B. 1969. Reliability for the researcher. *Alberta Journal of Educational Research* XV(2): 117-126.

Mayor, J.R., and Livermore, A.H. 1969. A process approach to elementary school science. *School Science and Mathematics* 69(5): 411-416.

Mechner, F. 1965. Science education and behavioral technology. In: *Teaching Machines and PROGRAMMED Learning II: Data and Direction*. Edited by R. Glaser. Washington: National Educational Association.

Micciche, F. 1970. *E.M.E. Hypothesis Machine Study Guide*. Bronxville, N.Y.: Educational Materials and Equipment. Mimeo.

__________, and Keany, M. 1969. Hypothesis machine. *Science Teacher* 36(4): 53.

Miller, S. A developmental study of hypotheses in attribute identification. Unpublished Ph.D. Dissertation, University of Colorado, 1971.

Mokosch, E. A process approach to science teaching. Unpublished Ph.D. Dissertation, University of Alberta, Spring 1969.

Munson, H.R. 1967. Evaluating the new science teaching. *Elementary School Journal* 68: 126-130.

National Society for the Study of Education (NSSE). 1960. The objectives of science teaching. In: *Rethinking Science Education*, 59th Yearbook of NSSE, Part I. Chicago: University of Chicago Press. pp. 39-46.

Nay, M.A. 1971. An inventory of processes in scientific inquiry. *Science Education* 55(2): 197-207.

Parker, J.S., and Rubin, L.J. 1966. *Process as Content: Curriculum Design and Application of Knowledge.* Chicago: Rand McNally.

Peel, E.A. 1965. Intellectual growth during adolescence. *Educational Review* XVII: 169-180.

Pella, M.O. 1961. The laboratory and science teaching. *Science Teacher* 28(5): 20-31.

Phillips, J.L. 1969. *The Origins of Intellect: Piaget's Theory.* San Francisco: Freeman.

Piaget, I. 1961. The genetic approach to the psychology of thought. *Journal of Educational Psychology* 52: 279-281.

__________. 1964. Cognitive development in children: Development and learning. *Journal of Research in Science Teaching* 2: 176-186.

__________. 1966. Psychology of intelligence and education. *Pennsylvania School Journal* 114: 318.

__________. 1970. *Science of Education and the Psychology of the Child.* New York: Orion Press.

__________. 1972. Intellectual evolution from adolescence to adulthood. *Human Development* 15: 1-12.

__________. 1973. *The Child and Reality: Problems of Genetic Psychology.* New York: Grossman.

Popham, W.J., and Sirotnik, K.A. 1973. *Educational Statistics: Use and Interpretation.* New York: Harper and Row.

Postman, L. 1951. Toward a general theory of cognition. In: *Social Psychology at the Crossroads.* Edited by J.H. Rahrer and M. Sherif. New York: Harper and Row.

Pulaski, M. 1971. *Understanding Piaget: An Introduction to Children's Cognitive Development.* New York: Harper and Row.

Quinn, M. Evaluation of a method for teaching hypothesis formation to sixth grade children. Unpublished Ph.D. Dissertation, University of Pennsylvania, 1971.

Ramsay, G.A., and Howe, R.W. 1969. Analysis of research on instructional procedures in secondary school science. *Science Teacher* 36: 62-66 (March); 72-81.

Reiner, W.B. 1966. Meeting the challenge of recent developments in science in assessing the objectives of modern instruction. *School Science and Mathematics* 66: 335-345.

Ripple, E., and Rockcastle, V.N. 1969. Piaget rediscovered. *Journal of Research in Science Teaching* 2: 168-252.

Robinson, J.T. 1965. Science teaching and the nature of science. *Journal of Research in Science Teaching* 3: 37-50.

Rogers, E.M. 1960. *Physics for the Inquiring Mind.* Princeton: Princeton University Press.

Rutherford, F. 1964. The role of inquiry in science teaching. *Journal of Research in Science Teaching* 2: 80-84.

Schools Council Publications. 1973. *With Objectives in Mind: Guide to Science 5-13.* London: Macdonald Educational.

Schwab, J. 1960. Enquiry, the science teacher and the educator. *Science Teacher* 27(6): 6-11.

__________, and Brandwein, P.F. 1962. *The Teaching of Science.* Cambridge, Mass.: Harvard University Press.

Schwebel, M., and Ralph, J. 1973. *Piaget in the Classroom.* New York: Basic Books.

Selberg, E.M., Neil, L.A., and Vessell, M.F. 1970. *Discovering Science in the Elementary School.* Don Mills: Addison-Wesley.

Stone, M., and Ausubel, D.P. 1969. The intersituational generality of formal thought. *The Journal of Genetic Psychology* 115: 169-180.

Tannenbaum, R.S. The development of the test of science processes. Unpublished Doctoral Dissertation, Teacher's College, Columbia University, 1969.

__________. 1971. The development of the test of science processes. *Journal of Research in Science Teaching* 8(2): 123-136.

Thier, H.D. 1970. *Teaching Elementary School Science.* Lexington, Mass.: D.C. Heath.

Thelen, H. 1960. The triumph of 'achievement' over inquiry in education. *The Elementary School Journal* 60(4): 190-197.

Thurber, W.A., and Collette, A.T. 1964. *Teaching Science in Today's Secondary Schools.* 2nd edition. Boston: Allyn and Bacon Inc.

Welch, W.M., and Pella, M.O. 1968. The development of an instrument for inventorying knowledge of the processes of science. *Journal of Research in Science Teaching* 5: 64-68.

Winer, B.J. 1962. *Statistical Principles in Experimental Design.* New York: McGraw-Hill.

Wood, D. 1960. *Test Construction.* Columbus: Merrill Books.

Woodburn, J.H. 1969. Hypotheses: Crossroads to success versus failure in science and science teaching. *School Science and Mathematics* 69(4): 333-337.

APPENDIX A

GENERAL SCIENCE TEST

JUNIOR HIGH GENERAL SCIENCE TEST

COMPILED BY

MORRIS TREASURE

SCIENCE EDUCATION CONSULTANT

DEPARTMENT OF EDUCATION

THIS IS A TEST BOOKLET. DO NOT WRITE OR MARK IN ANY WAY ON THE PAGES OF THIS BOOKLET.

Directions:

1. Print your name, school and grade on the answer sheet.

2. Indicate your age and sex in the proper space.

3. All answers in this test are to be machine scored. Each question has several suggested answers, one of which is the BEST answer. Select this BEST answer and record it on the separate answer sheet.

4. Please use an ORDINARY H.B. pencil to record your answers. Make sure your marks are heavy and black and that they do not extend beyond the guidelines. There should only be one choice marked for each question.

5. Do your best work but do not spend too much time on any one question.

180

1. In science the purpose of an experiment is to verify

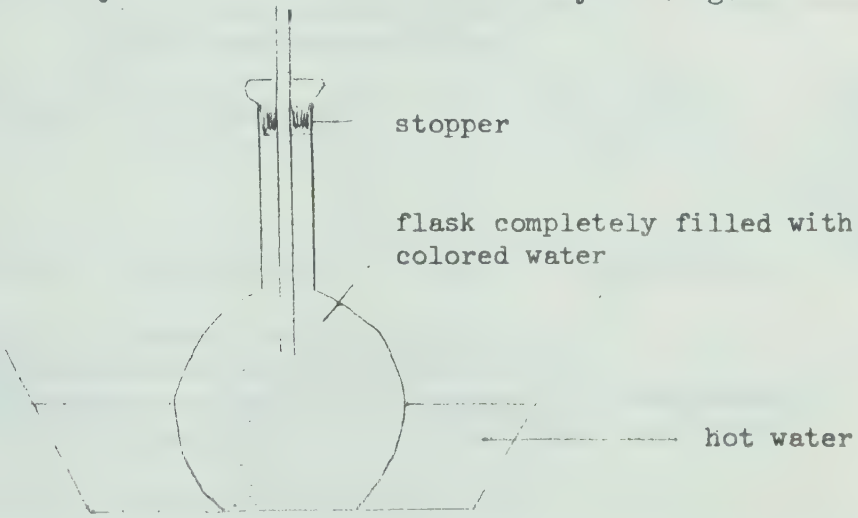
   A. An observation.
   B. A conclusion.
   * C. An hypothesis.
   D. A scientific law.

2. James was given 3 blocks of different metals which were labelled X, Y and Z. Each block had a volume of 20 cm^3. By using an equal-arm balance, he found that block X would just balance blocks Y and Z when placed together. However, block Y was not heavy enough to balance block Z. From this information, we know that the blocks in order of increasing density are:

   * A. Y, Z, X.
   B. X, Y, Z.
   C. Z, Y, X.
   D. Y, X, Z.

3. In the experiment below it was found that the water level dropped slightly and then rose rapidly. An explanation for this result would be that

   A. Evaporation causes cooling near the surface of the flask.
   * B. The flask expanded faster than the water.
   C. The water expanded faster than the flask.
   D. The density of the water is increased by heating.



4. A person trying to start a campfire uses shavings instead of larger pieces of wood because shavings

   A. Contain more energy per unit of weight.
   B. Catch fire at higher temperatures than larger pieces.
   C. Have a lower kindling temperature.
   * D. Have more surface area exposed to the air.

5. During a clear night various constellations appear and disappear as the earth rotates but the "Dippers" do not. This is because the stars forming the "Dippers"

   A. Do not move.
   B. Move at the same rate as the earth.
   C. Are located at or near the sky equator.
   * D. Are located at or near the north sky pole.

6. An iron bar was given in turn to each of three students. Each student measured the bar four times. The results were as follows:

TRIALS	1	2	3	4
Student A	34.000	34.001	34.001	34.002
Student B	34.003	34.003	34.004	34.005
Student C	34.005	34.006	34.006	34.006

(all readings are in inches)

What inference can be derived from the above table?

A. The students made inaccurate measures.
B. The readings should have been in centimeters, not inches.
C. The coefficient of expansion of iron is about 34.003.
* D. The heat from handling the bar caused it to expand.

7. Three experiements are set up. The observations are as follows:

Experiment I – Tiny particles suspended in water, when observed through a microscope, appear to be moving rapidly in short jerky motions.

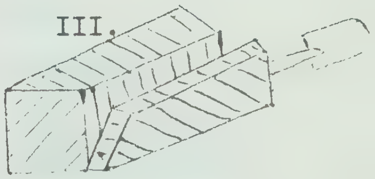
Experiment II – A piston when inserted into an air-filled cylinder does not drop to the bottom of the cylinder.

Experiment III – Some types of rocks when struck by a sharp blade will break off in very even sections.

I. View through microscope

II.

III.



Taken collectively, these might be evidence to suggest that

* A. Matter is composed of molecules.
B. Molecules are in constant random motion.
C. Temperature increases when molecules are compressed.
D. Light molecules travel faster than heavy molecules.

Use the following statements to answer items 8 - 12.

A. Most substances expand when heated and all substances have their characteristic expansion rates.

B. The quantity of heat required to raise the temperature a given amount varies between substances.

C. When matter changes state, there is a gain or loss of thermal energy without a change in the temperature of the substance.

D. Heat is transferred by radiation, conduction and convection.

8. A mercury thermometer indicates a change in temperature. 182

   * A. B. C. D.

9. A steam heating system transfers more heat than a hot water heating system.

   A. B. * C. D.

10. A burn from steam is more severe than a burn from boiling water, even though the temperatures are the same.

    A. B. * C. D

11. A thermostat may consist of two different metals bonded together.

    * A. B. C. D.

12. There is very little heat immediately above the surface of the moon, even though the surface may be quite hot.

    A. B. C. * D.

13. When we exhale through limewater, it turns milky, indicating the presence of carbon dioxide. What assumptions must be made by the experimenter if he is to accept this as a test for carbon dioxide?

    * A. Only carbon dioxide turns limewater milky.
    B. Only limewater turns milky when exposed to carbon dioxide.
    C. There are other gases present in exhaled air.
    D. There is more carbon dioxide in exhaled air than in inhaled air.

14. If a man puts his ear against a railroad track he can hear a train coming from a greater distance than when he is standing upright. What is the best explanation?

    A. Hearing is more sensitive when the head is closer to the ground.

    * B. The more dense a substance, the faster it will conduct vibration.

    C. Air is a better conductor of sound than steel.

    D. Air is more dense near the ground and therefore a better conductor of sound.

15. A student reported to her science class that water in the form of ice evaporates at temperatures below zero. She said her mother's clothes dry on the outside line when it is -30°F.

    What is the best conclusion?

    A. The water dripped off the clothes before they froze.
    * B. Some solids may evaporate like liquids.
    C. Water cannot evaporate at temperatures below freezing.
    D. The student made observational errors.

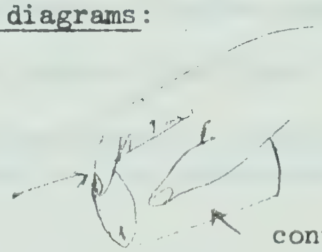
16. A student took two tiny pieces of paper and suspended each by a piece of silk thread. He touched both papers to remove any electrical charges which might have been present. Then he rubbed a piece of ebonite rod on fur so as to charge it negatively. The ebonite was then placed on one piece of paper and removed. Immediately, the papers flew together and then flew apart. A possible suggestion to explain this might be that

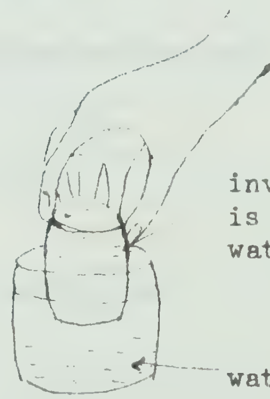
   A. Like charges attract and unlike charges repel.
   B. Like charges repel and unlike charges attract.
 * C. A charged object attracts an uncharged object and like charges repel.
   D. A charged object will attract any object and an uncharged object will not attract any object.

17. A student performed an experiment to prove that plants give off moisture. A plant was completely covered by a large glass container as shown below. The apparatus was left outside for a day where the temperature was 45^{o}F. He noticed moisture had formed on the inside of the glass container. He then moved the apparatus into a room where the temperature was 75^{o}F. After a few hours the moisture had disappeared. The probable cause was that the moisture

 * A. Was absorbed by the air inside the glass.
   B. Was absorbed by the plant.
   C. Changed to heat energy.
   D. Was absorbed by the soil.

18. <u>Study these diagrams</u>:





Water does not run out of either container because

   A. Air weighs less than water.
   B. Suction holds the water in the containers.
   C. There is a vacuum above the water
 * D. The pressures are balanced.

19. Large vegetable storage houses (root houses) used to be constructed with no outside source of heat used in the winter. Often there was a large tank of water in the middle. What is the best reason for placing the tank of water in the storage bin?

   A. The water keeps the vegetables crisp.
 * B. The water helps to maintain a constant temperature in winter and summer.
   C. The vegetables absorb the water vapor which slows down the rotting process.
   D. The water vapor in the air would increase the pressure in the bin and keep the cold air out.

20. A 500 ml. flask was half-filled with water, closed with a solid rubber stopper and placed on a tripod stand over a Bunsen burner. As the water was being heated, the rubber stopper was blown out of the mouth of the flask. The most logical explanation is that

    A. Heat caused an increase in the size of the molecules in the flask.

    B. Heat caused an increase in the number of molecules in the flask.

    C. The water expanded and forced the air to rise pushing out the stopper.

    * D. Gas molecules moved faster causing more numerous collisions with the stopper.

21. A student removed all the air he could from a pop bottle by sucking on the mouth of it. He then placed the mouth of the bottle under water to see how much air he had removed. He carefully measured and found that the water rose only 3 inches up the bottle. He repeated his experiment, but used a different liquid the second time. This time, the liquid rose over 4 inches up the bottle. Several of his classmates repeated his experiment and got similar results. Which of the following is the best explanation for this.

    * A. The second liquid was less dense than the first.

    B. The practice of removing the air the first time enabled the students to remove more air the second time.

    C. The second liquid must have been warmer than the first.

    D. The atmospheric pressure must have increased between the two parts of the experiment.

For Items 22 - 27:

Mark A if statement I <u>causes</u> statement II
Mark B if statement II <u>causes</u> statement I
Mark C if both statements are the <u>result</u> of the <u>same phenomena</u>
Mark D if the two statements are <u>not related</u> in any way

22. I. During the past 50 years extensive land clearing has been done in Alberta.

II. There has been an increased water flow in streams and rivers in Alberta.

* A. B. C. D.

23. I. Mountain flowers are small and have a short growth cycle.

II. Arctic flowers are small and have a short growth cycle.

A. B. * C. D.

24. I. A fire destroyed a mature forest. New growth began to appear.

II. Two years later the deer population had greatly increased in the area.

* A. B. C. D.

25. I. The lights in the room became dimmer.

II. A black space appeared around the picture on the T.V.

A. B. * C. D.

26. I. There is a rapid decrease in the number of game birds.

II. Farmers have eliminated brush in order to increase crop acreage.

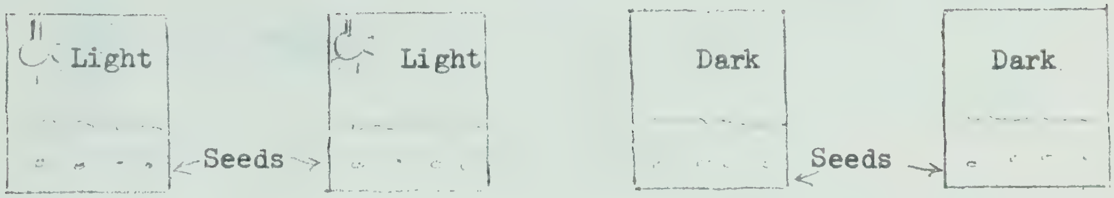
A. * B. C. D.

27. I. The province of Alberta paid a bounty on wolves and coyotes, thus reducing their number.

II. In five years time it was found that deer in the province were not as healthy as before.

* A. B. C. D.

28.



Dry Soil A   Moist Soil B   Dry Soil C   Moist Soil D

The experiment was set up and operated for one month.

If the student's objective was to study the effects of light on germination, the box(es) which acted as control(s) would be

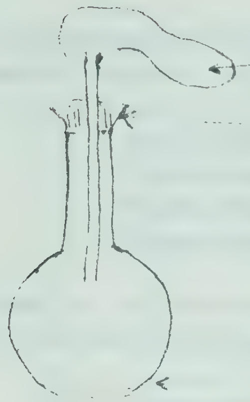
A. A.
B. A and B.
C. C.
* D. C and D.

29. If the student's objective was to study the effects of moisture on germination, the box(es) which acted as control(s) would be

A. A.
* B. A and C.
C. D.
D. B and D.

30. A condition which was not controlled is

* A. Temperature.
B. Moisture.
C. Light.
D. Time.



After a student placed baking soda in the flask, he poured in dilute vinegar and quickly stoppered the flask with a tight fitting, one holed, rubber stopper carrying a snug fitting glass tube topped by a deflated balloon. The balloon quickly inflated.

31. The main reason the balloon inflated is that

    A. Pressure outside the flask increased.
    * B. Pressure inside the flask increased.
    C. Air in the flask expanded as a result of cooling.
    D. Pressure inside the flask decreased.

32. The type of reaction which occurs in the flask is most clearly related to

    A. The propulsion of a jet engine.
    * B. The action of a soda-acid fire extinguisher.
    C. Food digestion in the stomach.
    D. Photosynthesis in green plants.

33. If the stopper is removed from the flask and a lighted match is inserted, the most likely result would be that

    * A. The match will go out.
    B. The gas in the flask will ignite.
    C. The match will flare up.
    D. There will be no noticeable change.

34. After a jar of cold tap water sits in a warm room for some time, bubbles collect on the edges of the glass. This can best be explained by the fact that

    A. More air has dissolved in the water.
    B. Some of the water has vaporized and formed bubbles.
    C. Air is more soluble in warm water than it is in cold water.
    * D. Air is more soluble in cold water than it is in warm water.

35. A boy in central Alberta using a magnetic compass to determine direction found that the compass needle pointed considerably to the east of a street he knew to be running due north. The situation results from the fact that the

    A. Earth's geographic north pole is not due north of central Alberta.
    B. North star no longer lies over the North Pole.
    C. Earth's geographic north pole was arbitrarily located.
    * D. Earth's magnetic pole is not due north of central Alberta.

36. A lid that is hard to remove from a bottle of ketchup may be loosened by immersing the top of the bottle in hot water because 188

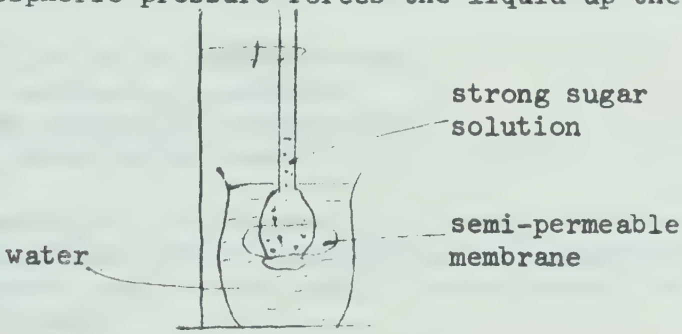
A. Hot water dissolves the dried ketchup.
* B. Solids expand at different rates.
C. Glass expands at an uneven rate.
D. Metals are good conductors of heat.

37. Fossils of the earliest forms of animal life are found in the deepest layers of the earth's crust. The best explanation for this fact is that

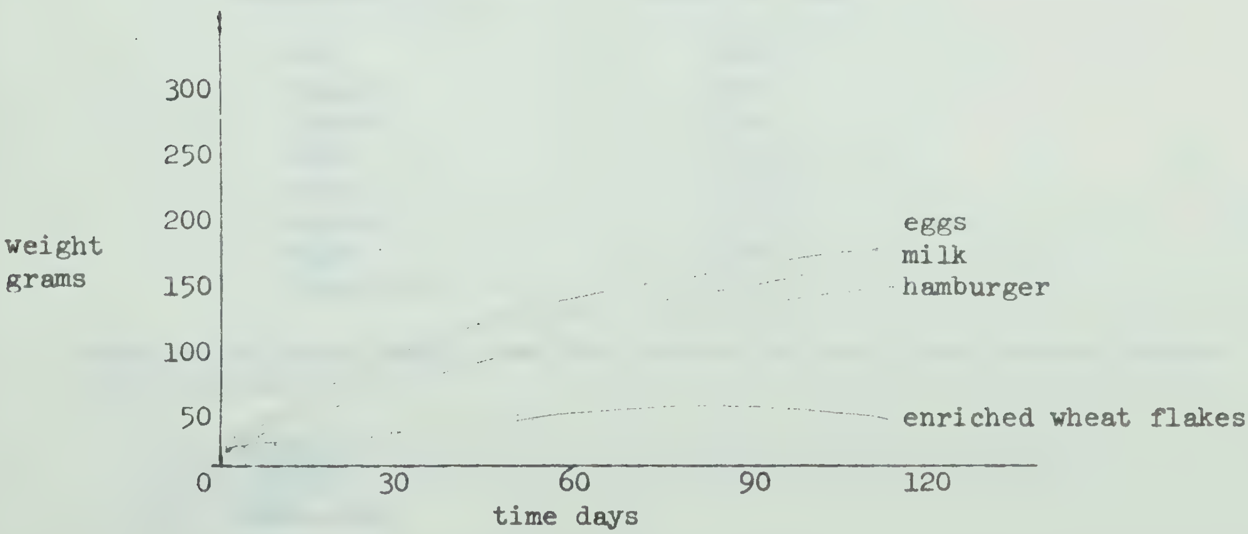
A. These are fossils of small simple forms of life.
B. The deepest layers of rocks are undisturbed by earth movement.
C. There are relatively few fossils of later forms of life.
* D. The deepest layers were formed before the upper layers.

38. What conclusions would you draw from the diagram?

* A. The movement through the membrane is toward the more concentrated solution.
B. The movement through the membrane is toward the less concentrated solution.
C. The larger surface area of the water allows more evaporation.
D. Atmospheric pressure forces the liquid up the tube.



39. RAT GROWTH ON DIFFERENT FOODS



The graph shows that

A. Enriched wheat flakes lead to steady growth.
* B. Rats thrive on a diet of eggs.
C. Milk is a good dietary source of protein for rats.
D. Hamburger results in steady growth after 200 grams.

40. Use the four statements below to answer the following questions.

I. Two suspended ping-pong balls move together as a rapidly moving air stream is passed between them.

II. As wind rushes by the open door of a room, papers in the room seem to be sucked out by the wind.

III. A piece of paper drooped over the edge of a desk rises as an air stream is passed over the paper parallel to the desk.

IV. A ping-pong ball, floating freely, remains suspended in a stream of air directed vertically upwards.

From the above statements, one might conclude that

A. Light objects are easily moved from their positions.
B. A rapidly moving stream of air exerts a considerable amount of force.
* C. Speeding up of air causes a decrease in its pressure.
D. A rapidly moving air stream, causes light objects to move unpredictably.

41. An application of the principle common to the four statements is

* A. The lift on an airplane wing.
B. The tacking of a sailboat.
C. A fan forcing air into motion.
D. Pop rising in a straw.

42. The heat capacity of a substance is the amount of heat energy necessary to raise the temperature of one gram of the substance through one centigrade degree. Specific heat is the ratio of heat capacity of a substance as compared with water.

Substance	Specific heat
Air	0.24
Alcohol	0.66
Copper	0.09
Iron	0.119
Lead	0.031
Mercury	0.033
Water	1.00
Petroleum	0.51

Which fluid would act as the best carrier of heat in a heating system in a factory?

A. Alcohol.
B. Mercury.
C. Petroleum.
* D. Water.

43. The coefficient of linear expansion for five metals is given in the table below: 190

Aluminum	.000023
Brass	.000019
Copper	.000017
Iron	.000012
Platinum	.000009

Which of the following pairs, used in a bimetal strip, will react most, to small change in temperature?

A. Brass and Iron.
* B. Aluminum and Platinum.
C. Copper and Iron.
D. Brass and Platinum.

44. A huge boulder was found in the middle of a farmer's field. The best explanation as to how it got there would be that.

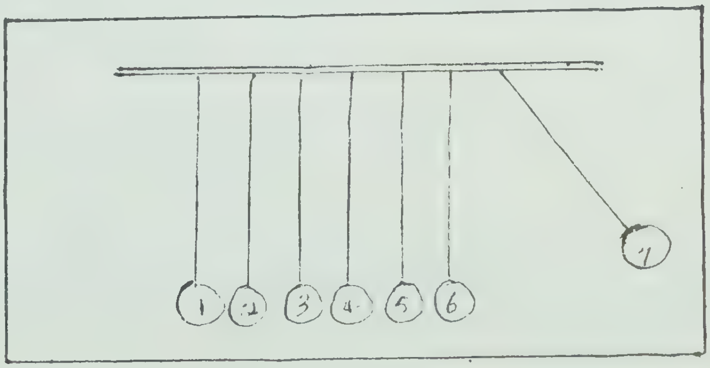
A. The rock fell from a mountain.
* B. Glacial movement carried the rock.
C. The rock was a meteorite.
D. Volcanic action had taken place.

45. Use the following three statements to answer question

I. Long, deep scratches in bed rock run in a north-south direction.
II. Extensive areas of gravel and sand hills can be found in Alberta.
III. Rocks are present that are not normally found in an area.

Taken collectively, these observations might suggest evidence of

A. Ancient sea shores.
B. The carboniforous era.
* C. The ice-age.
D. Extinct volcanos.



46. Set up seven pendulums all of equal mass as shown in the diagram. Add energy to pendulum 7 by moving it to the right and then letting it go. This energy would cause

A. All pendulums to move to the left.
B. A sudden jar on pendulums 1 through 6 and set all of them swinging back and forth.
* C. Pendulum 1 to move to the left approximately the same distance as pendulum 7 was moved to the right.
D. All pendulums to sway to the left and return to the hanging position.

47. The Northern Snowy Owl of Canada feeds in summer on lemming, which are small mouse-like creatures inhabiting the Arctic Barrens. Lemming are also the staple food of the Arctic fox. About every four years, the supply of lemming falls and the Snowy Owl must fly south to catch mice and rats. When the Eskimos of the Arctic region note a migration or movement of the owls out of the Arctic regions, they would predict that

A. The winter will be long and cold.
B. Southerly regions will have an abundance of mice and rats.
* C. Arctic fox will be scarce in the Arctic regions.
D. Snowy Owls will soon become extinct.

The data below represents the design and observations of an experiment. Use the information to answer the questions which follow.

Two pieces of brass, one with a mass of 500 grams and the other with a mass of 1000 grams, were placed in a large bunsen flame along with a piece of copper which also had a mass of 1000 grams. The three were heated to a temperature of 500°C. the 500 gram piece of brass was placed into calorimeter A. the 1000 gram piece of brass was placed into calorimeter B, and the 1000 gram piece of copper was placed into calorimeter C. Each calorimeter was identical and contained 3000 grams of water at 5°C. In 2.5 minutes the temperature in calorimeter B was 43°C and the temperature of the water in calorimeter C was 36°C.

	A	B	C
	500 gm of Brass at 500°C	1000 gm of Brass at 500°C	1000 gm of Copper at 500°C
	5°C 3000 gm H_2O	5°C 3000 gm H_2O	5°C 3000 gm H_2O
Final temperature=	24°C	43°C	36°C
Temperature change=	19°C	38°C	31°C

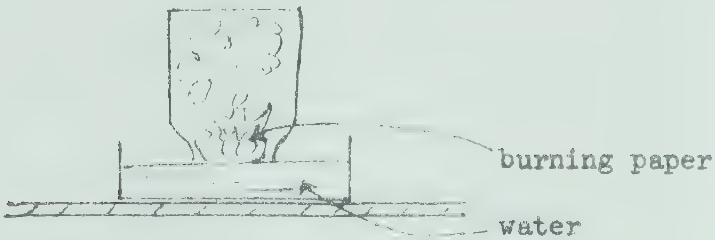
48. Two main inferences concerning heat can be drawn from the experiment. Which of the following would represent one of those inferences?

* * A. The heat capacity of an object depends upon the material from which the object is made.
* B. Water has a greater heat capacity than that of copper or brass.
* C. The mass of an object has no effect upon the amount of heat which it contains.
* D. Copper loses heat more quickly than brass.

49. This experiment actually deals with two problems concerning heat. Which of the following would represent one of those problems?

* A. Does heat travel from a metal to a liquid?
* B. Does heat flow from a hot material to a cooler material?
* * C. Does the amount of heat contained in an object depend upon its mass?
* D. Is the capacity of water greater than that of brass or copper?

50. In an attempt to show that the atmosphere is about 1/5 oxygen, a demonstration was set up as shown below:



It was assumed that the flame would remove the oxygen from inside the jar and the difference in pressure would then force water into the jar until about 1/5 of the space was occupied. The water, however, entered to fill much less than the expected amount. The demonstrator failed to take into account that

A. The air is more than 1/5 oxygen.
B. Nitrogen will also burn.
* C. Gases expand when heated.
D. The water is heavier than the air inside.

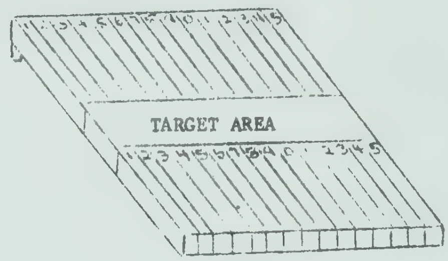
APPENDIX B

INFERENCE TEST

INFERENCE TEST

Morris Treasure

This series of exercises is based on a piece of apparatus that uses small ball bearings in a series of 15 channels. The channels are interrupted in the middle by an open space called the target area. Various objects can be placed in the target area and these have different effects on the ball bearings as they roll down the channel.



The objects used in this exercise are:

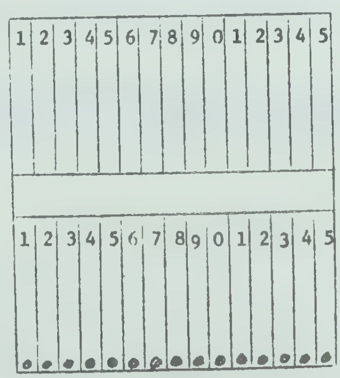
1. cups –  of different sizes
2. blocks –  of different sizes
3. slopes –  or  of different sizes

On the basis of the distribution of the balls, one of which has been rolled down each numbered channel, what are the positions, sizes and shapes of the object in the target area?

Draw in the object that you think has caused the observed distribution. Be as careful as you can making the object large enough to cause the distribution.

For example:

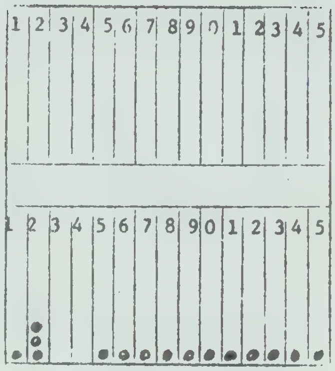
This distribution would indicate that there is nothing in the target area. On the answer sheet you would indicate by either leaving the space blank or writing the word "nothing" in the target area.

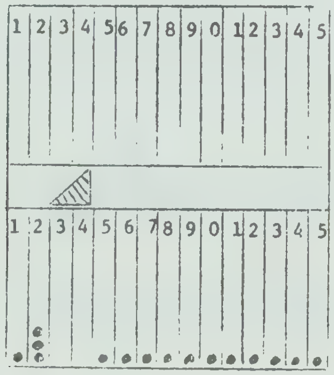


June, 1972

Example 3:

In this case two balls have been deflected from their expected destination. Therefore some object has interfered with their path. On the answer sheet you should draw in a slope, two channels wide, in the target area.





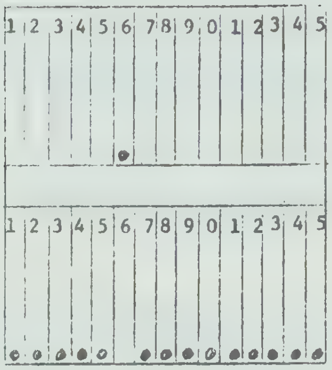
If a ball is not shown, "perched" or "deflected" it is probably "cupped" or captured by a cup-shaped object. Show the missing balls in the cupped object as you draw it.

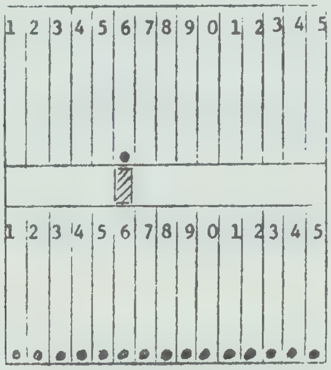
There may also be a combination of blocks, cups and slopes of various sizes in the target area.

Now begin the exercises and continue until the teacher tells you to stop. If you are finished before you are told to stop please indicate the length of time it took you to finish. Note the time now ________(time) and begin the first exercise.

Example 2:

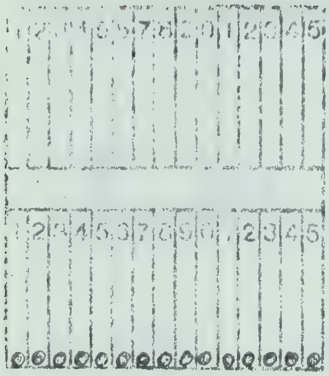
In this case, there is one ball "perched" in channel 6. Therefore something has blocked the ball's progress. On the answer sheet you should draw in a block, one channel wide, in the target area





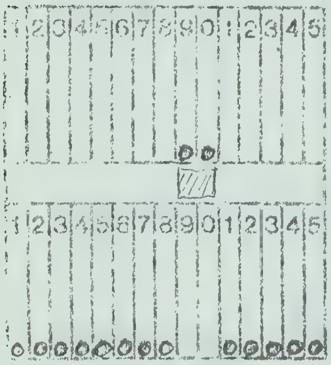
INFERENCE TEST

Page 1

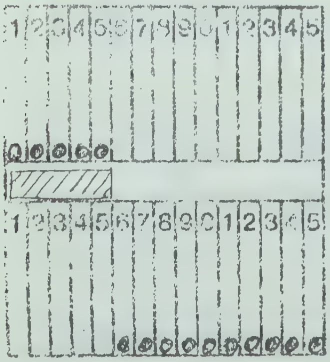


1

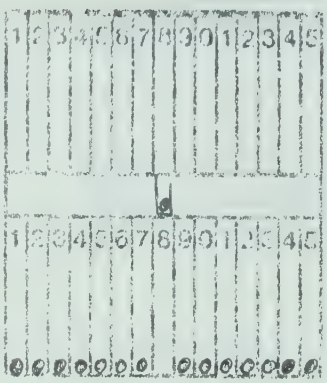
2



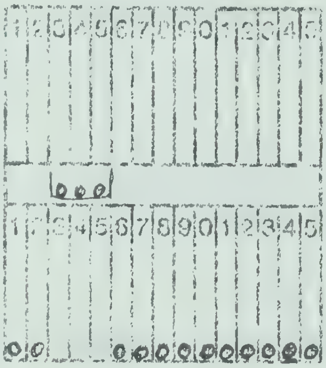
3



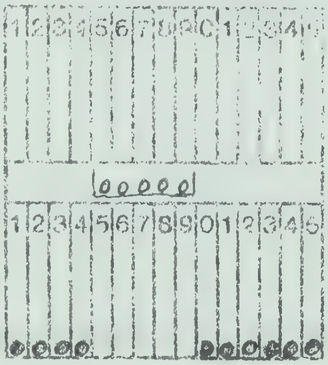
4



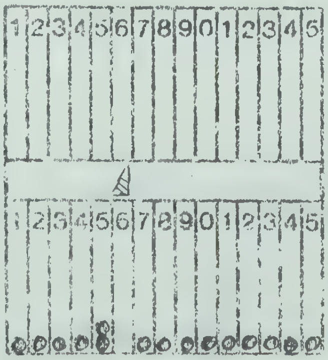
5



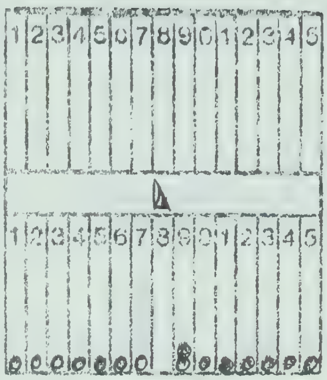
6



7

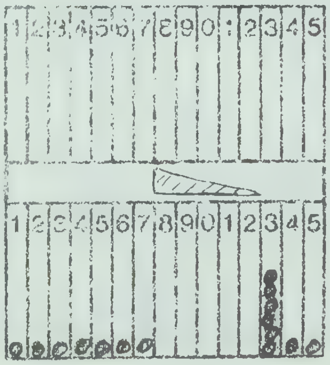


8

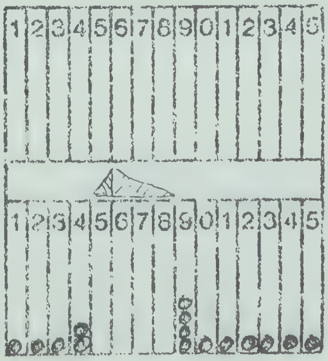


9

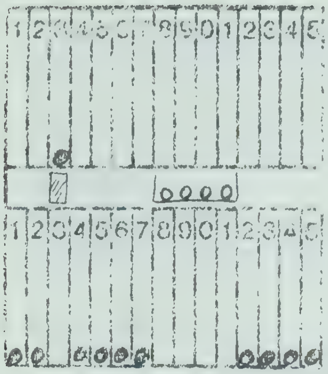
10



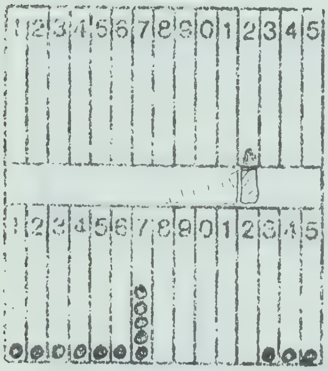
11



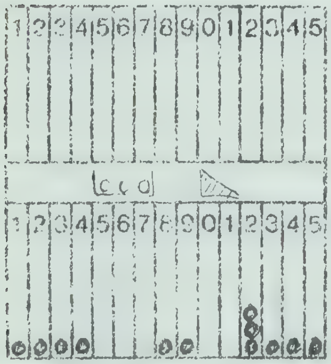
12



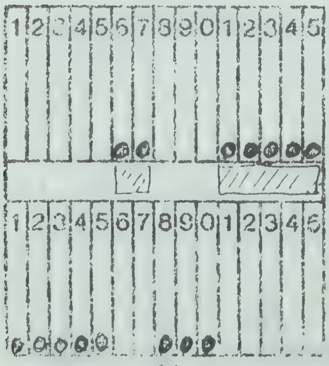
13



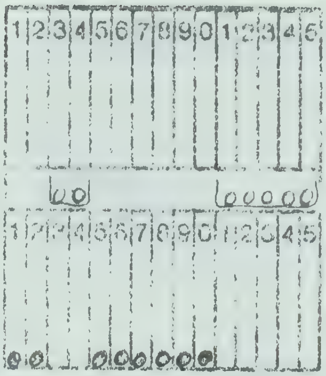
14



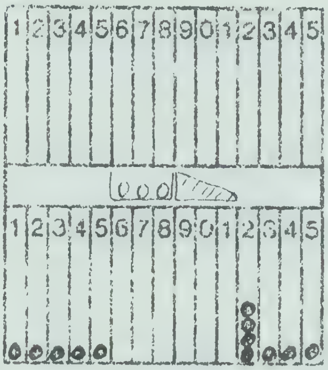
15



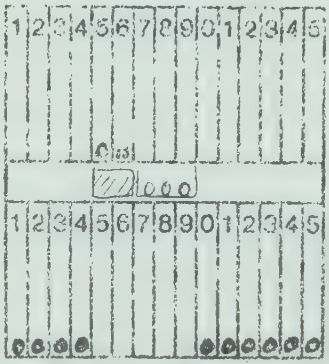
16



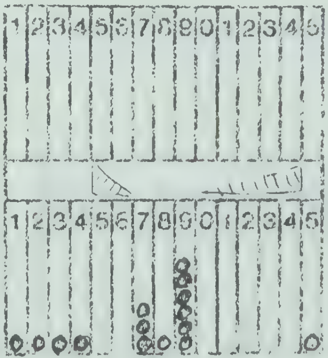
17



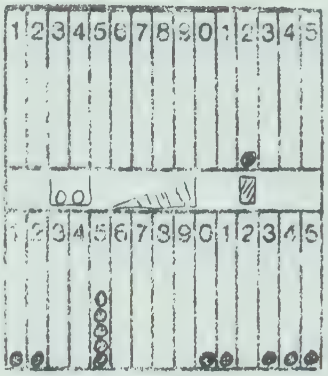
18



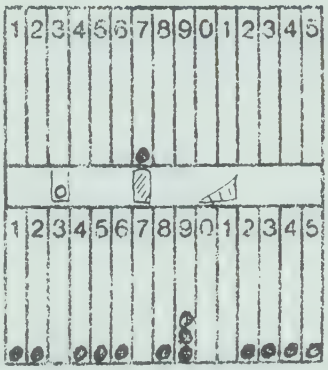
19



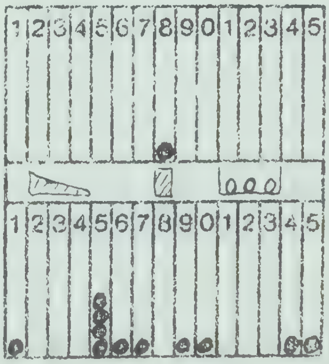
20



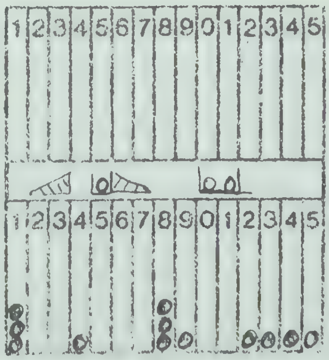
21



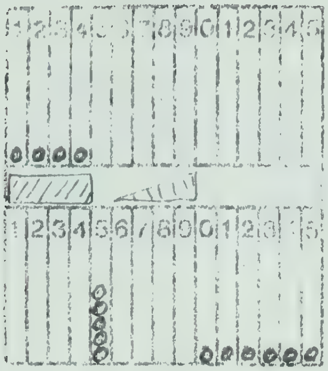
22



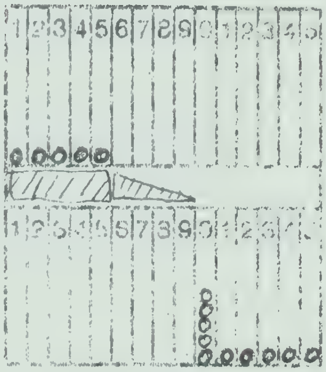
23



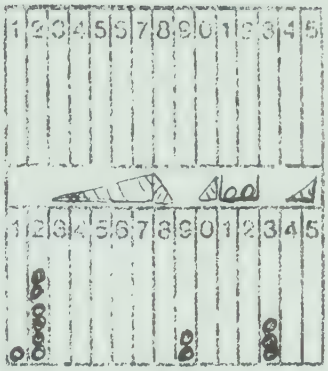
24



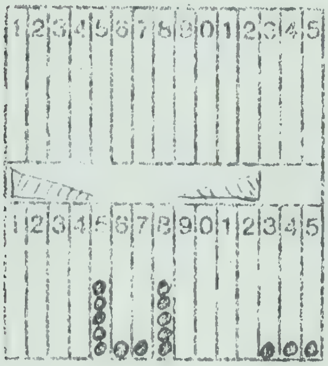
25



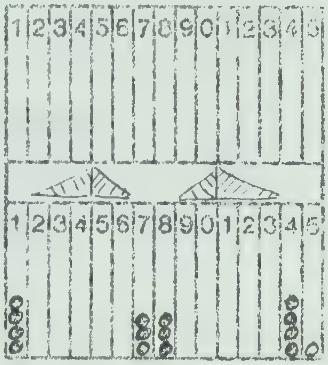
29



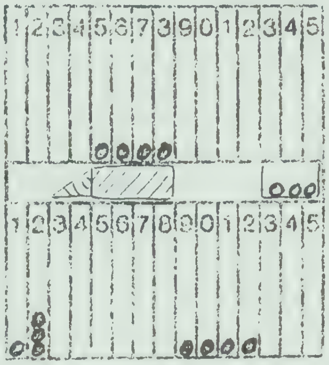
33



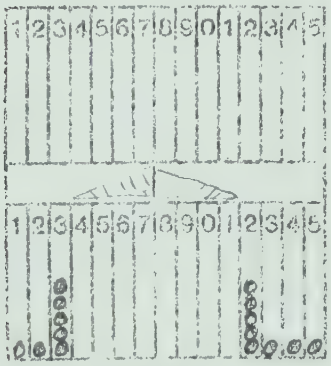
26



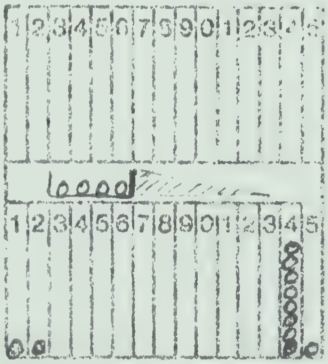
30



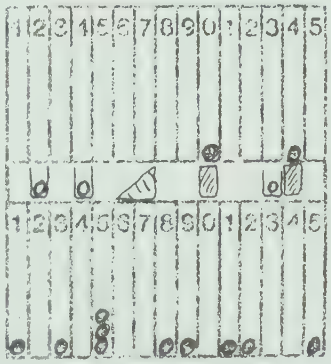
34



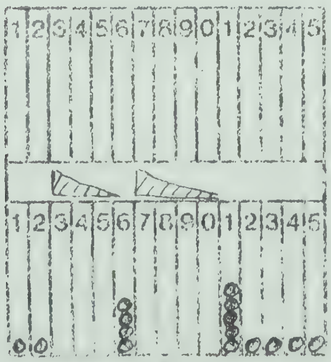
27



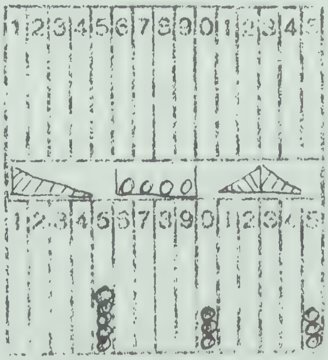
31



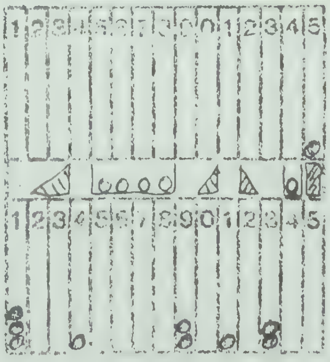
35



28



32



36

APPENDIX C

HYPOTHESIS TEST

HYPOTHESIS TEST

Morris Treasure

This is a test of an individual's ability to develop an hypothesis based on an inference from given data. This hypothesis is then subjected to a test and modified in the light of further inferences and data.

The given data is in the form of the distribution of ball bearings in a piece of apparatus with 4 channels interrupted by an open space called the target area.

In this series of exercises various objects are placed in the target area. The ball bearings are then dropped, one in each channel and the resulting distribution forms the given data. To lessen confusion and reduce the number of ball bearings the apparatus has had channels 1 and 4 reduced and since there will never be balls in 2 and 3 when there is a target in place, these channels have been modified so the apparatus now looks like:

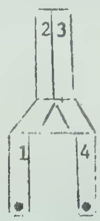
From the ball distribution, an inference is made about the nature of the object in the target area. On the basis of a number of inferences an hypothesis is formulated. A prediction is then made about the ball distribution for another orientation of the object. The ball distribution is shown for the object in four orientations.

The objects are stylized capital letters of the alphabet. Each letter is shown in four orientations--normal, rotated 90° clockwise (on its side), rotated 180° clockwise, (upside down), rotated 270° clockwise (on other side). The fifth position is a return to the normal, upright orientation. Two balls are dropped one in each channel with the object in each position - no balls are removed.

The inferences that are to be made refer only to whether the object in the channel is a block, a slope or a cup.

<u>Example 1.</u> The letter has been placed in the target area and two balls are dropped each time.

Position 1



Inference

#2	#3
block	block
slope	slope
cup	cup

Position 2 

Inference:

#2	#3
block	block
slope	slope
cup	cup

Position 3 

#2	#3
block	block
slope	slope
cup	cup

Position 4 

#2	#3
block	block
slope	slope
cup	cup

* note - balls appeared in both channels 1 & 4 .'. the balls that were in a cup in position 3 were dropped on rotation.

Position 5 

Hypothesis:

The letter must be A since position 1 & 5 are the normal orientation.

If two more balls are dropped the distribution would be

Example 2. Data Given

Position 1 

Inference

#2	#3
block	block
slope	slope
cup	cup

What you can't see

Position 2 

#2	#3
block	block
slope	slope
cup	cup

*The original two balls were dropped in channel 4 as the object rotated

Position 3

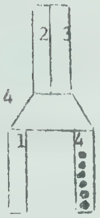
Inference

#2	#3
block	block
slope	slope
cup	cup

*The balls from position two are still hidden. They will not drop until the letter is again turned.

3

Position 4

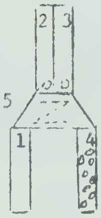


#2	#3
block	block
slope	slope
cup	cup

*The balls from both position two and three have been dropped in channel 4.

4

Position 5



Hypothesis:

The letter is ______.

5

If two more balls are dropped what is the distribution to be seen?

Instructions

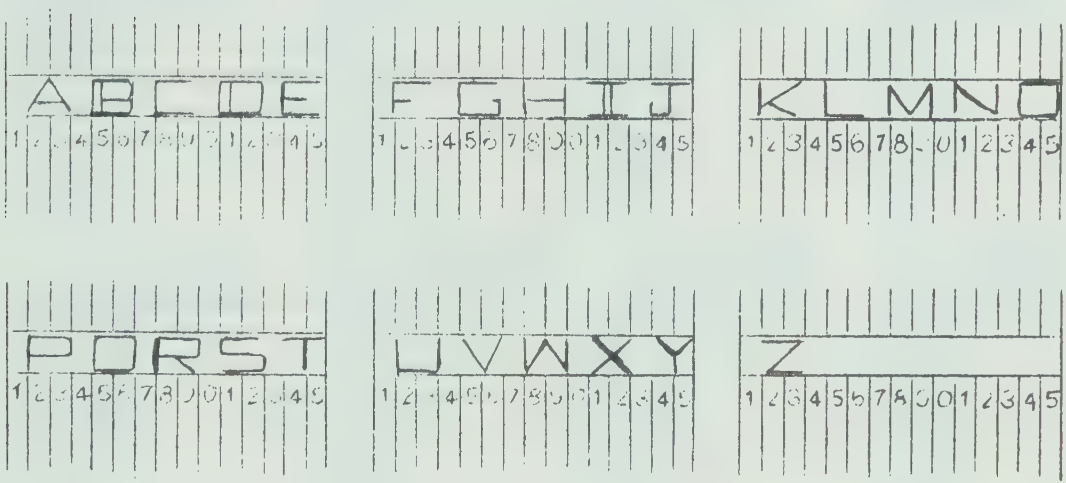
On the next pages you will be given the ball distributions for a number of objects. Each object is presented in four consecutive positions.

e.g. 

Each time the object is rotated two more balls are dropped. You are to determine whether the distribution would infer a block, a slope or a cup in each of channel 2 and 3. Circle the appropriate term.

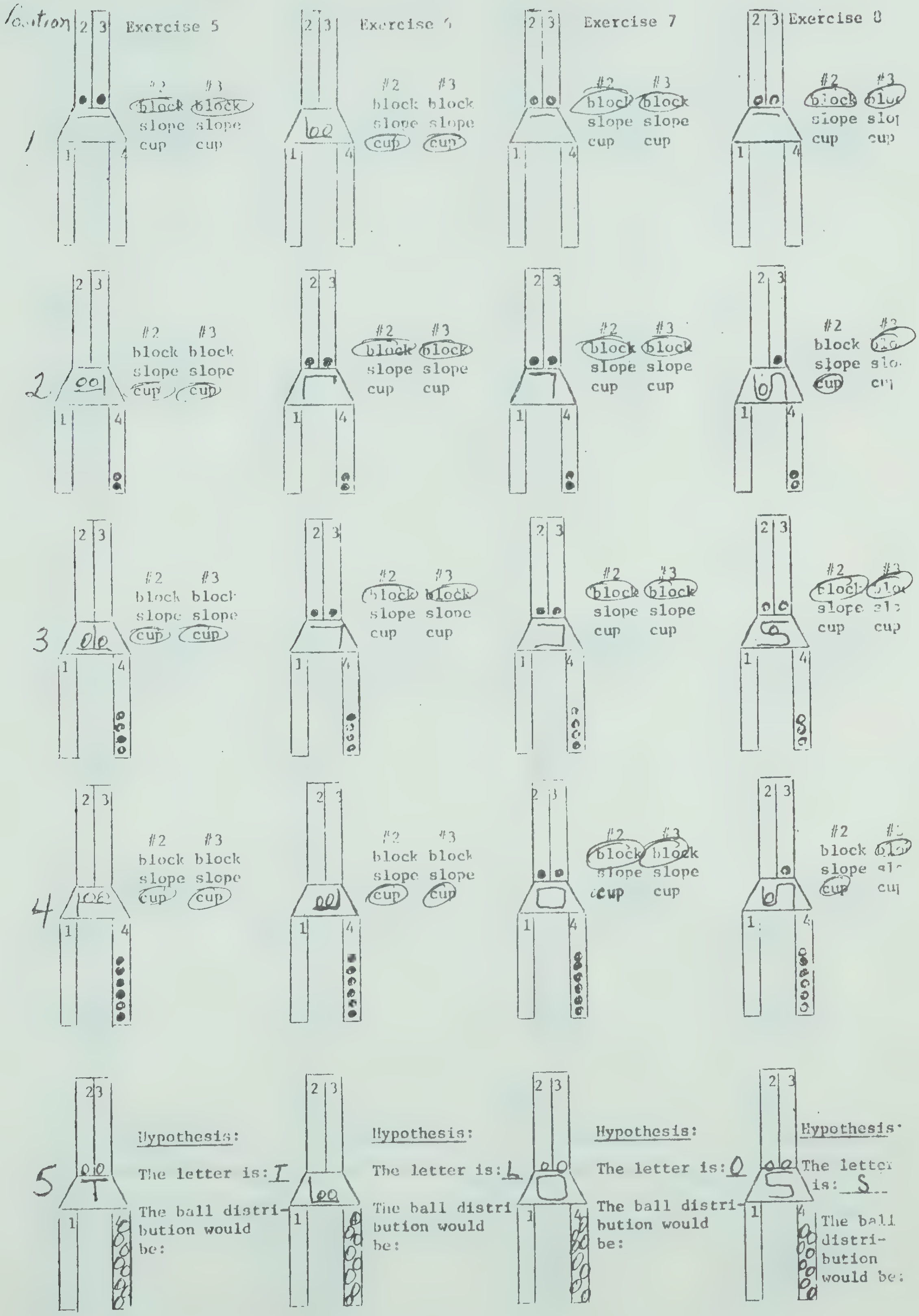
In the fifth position show the letter you have guessed it to be, and the appropriate ball distribution after two more balls have been dropped. (Show a total of 10 balls).

The stylized alphabet used is:

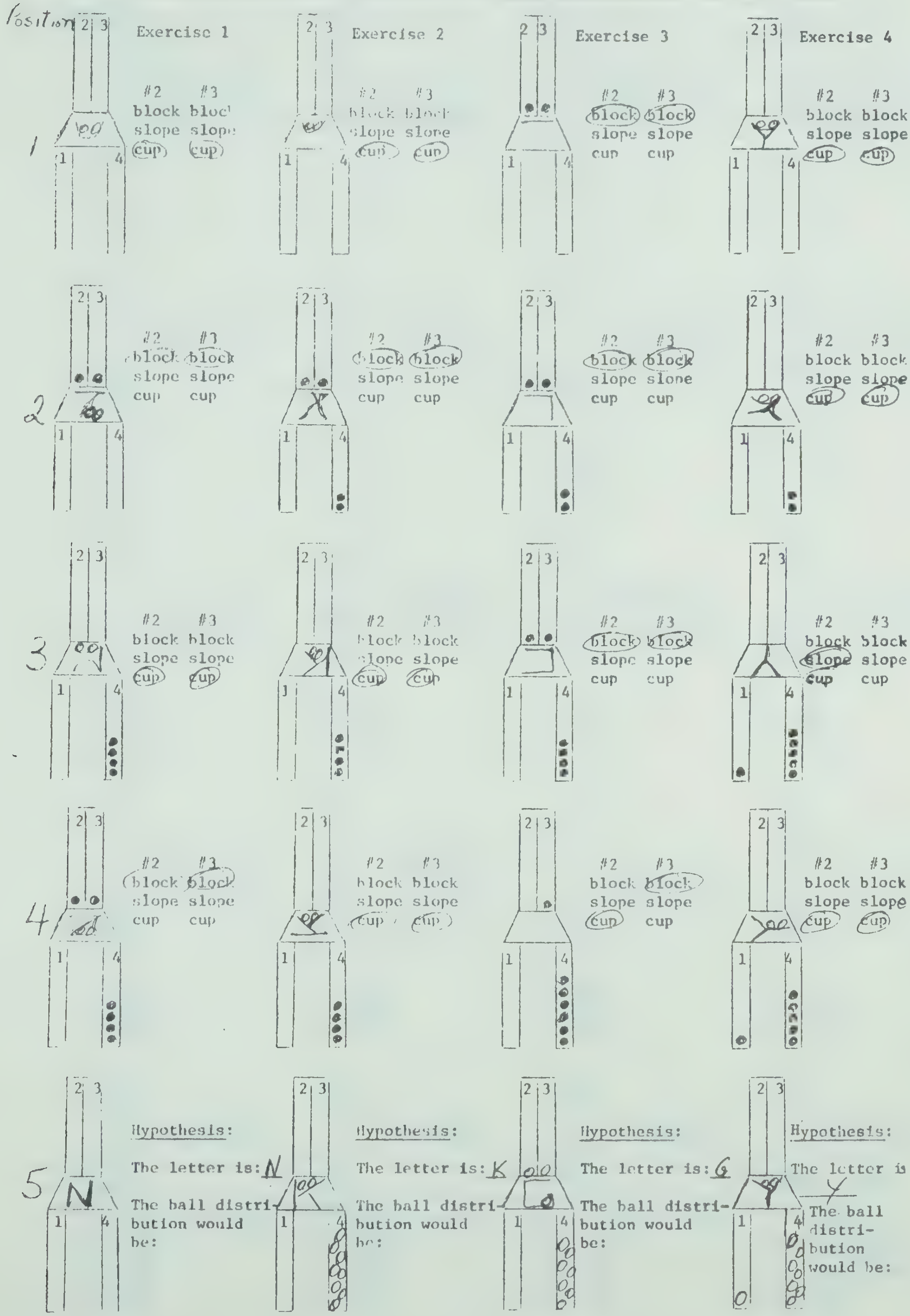


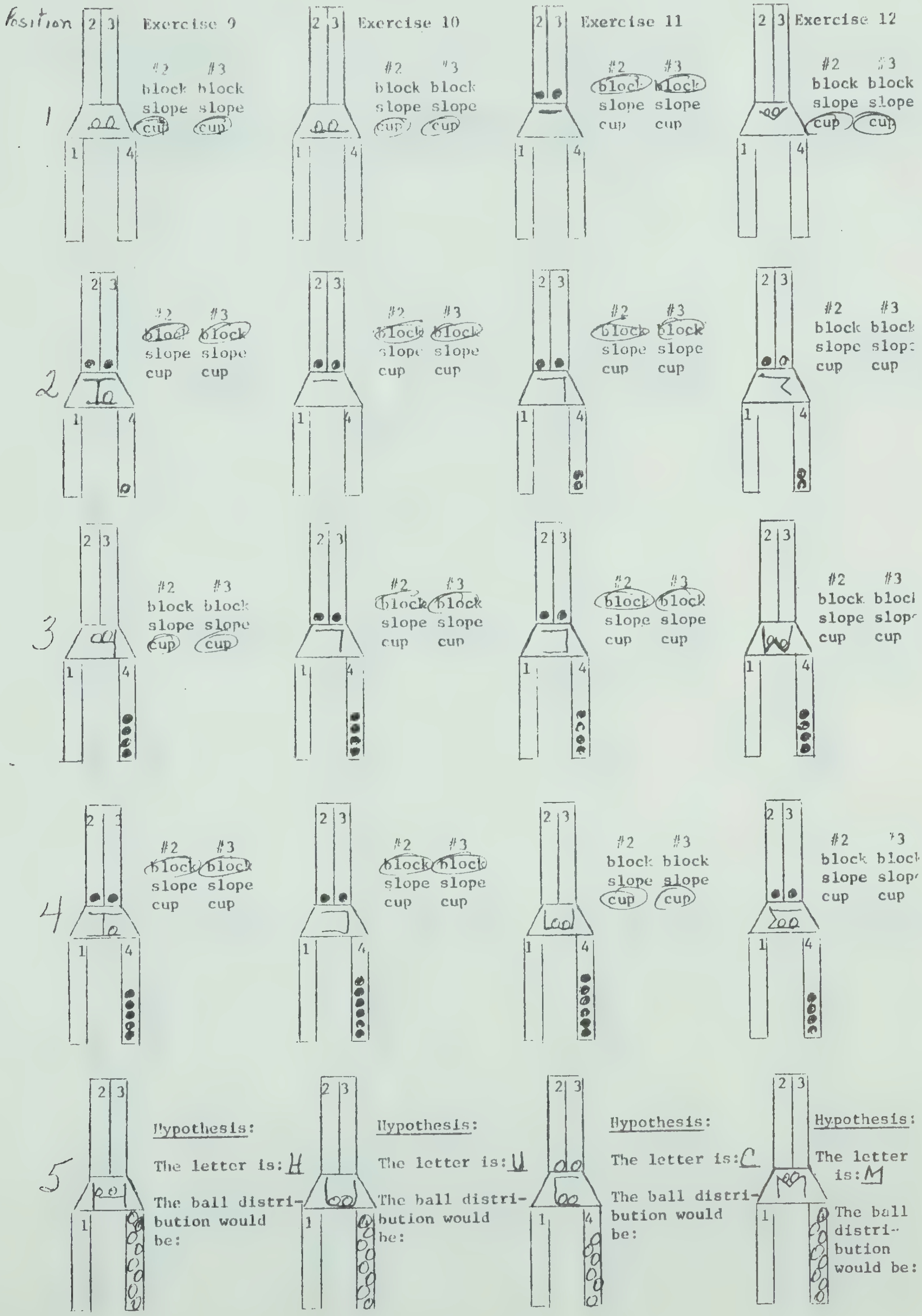
Note that each letter is two channels wide and fits the target area exactly.

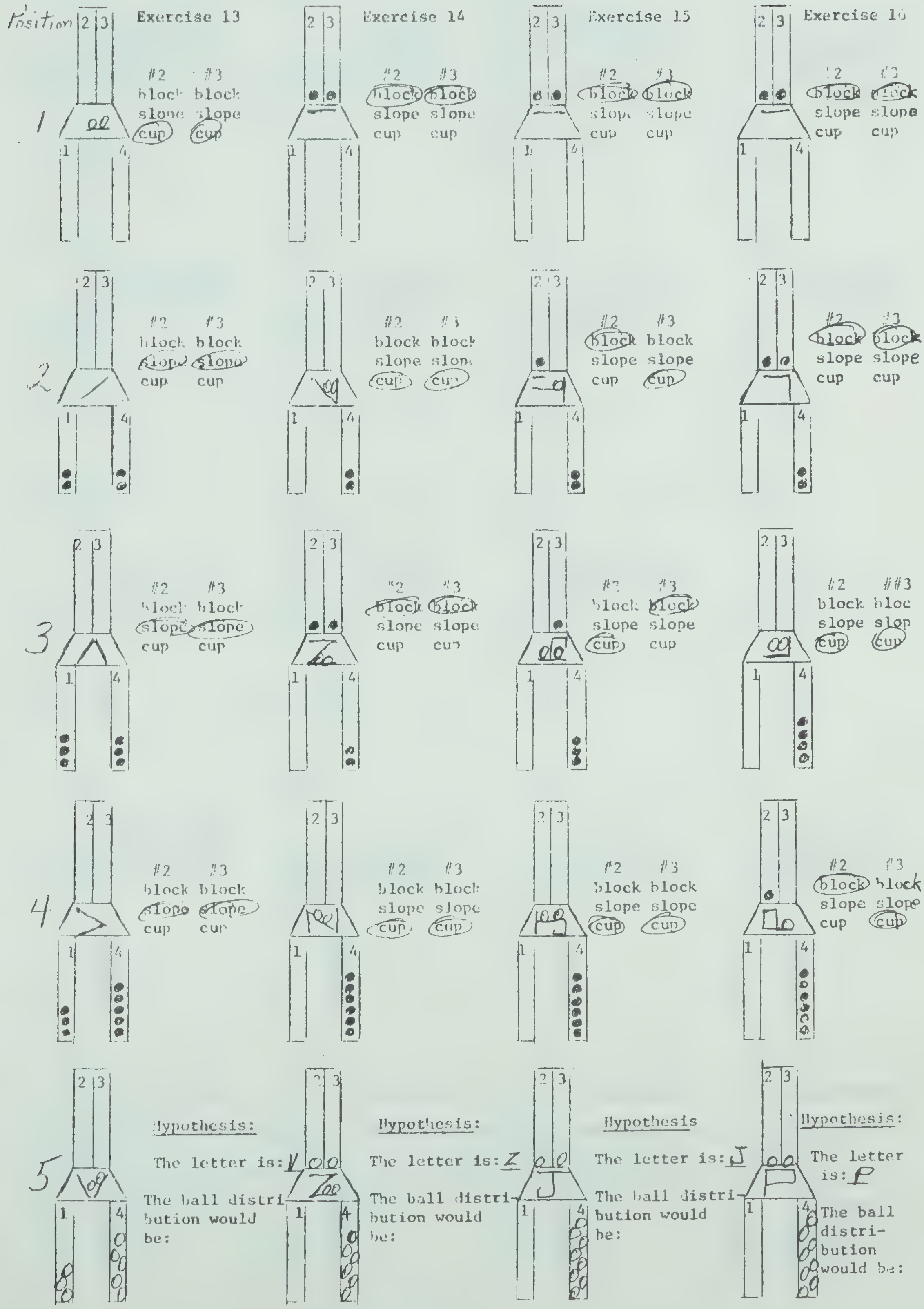
You have 20 minutes in which to complete the exercise. If you finish before the 20 minutes has elapsed please indicate the length of time that you took.


Position
Exercise 5
Exercise 6
Exercise 7
Exercise 8
1
2
3
4
5
#2 #3
block block
slope slope
cup cup
Hypothesis:
The letter is: I
The ball distribution would be:
Hypothesis:
The letter is: L
The ball distribution would be:
Hypothesis:
The letter is: O
The ball distribution would be:
Hypothesis:
The letter is: S
The ball distribution would be:

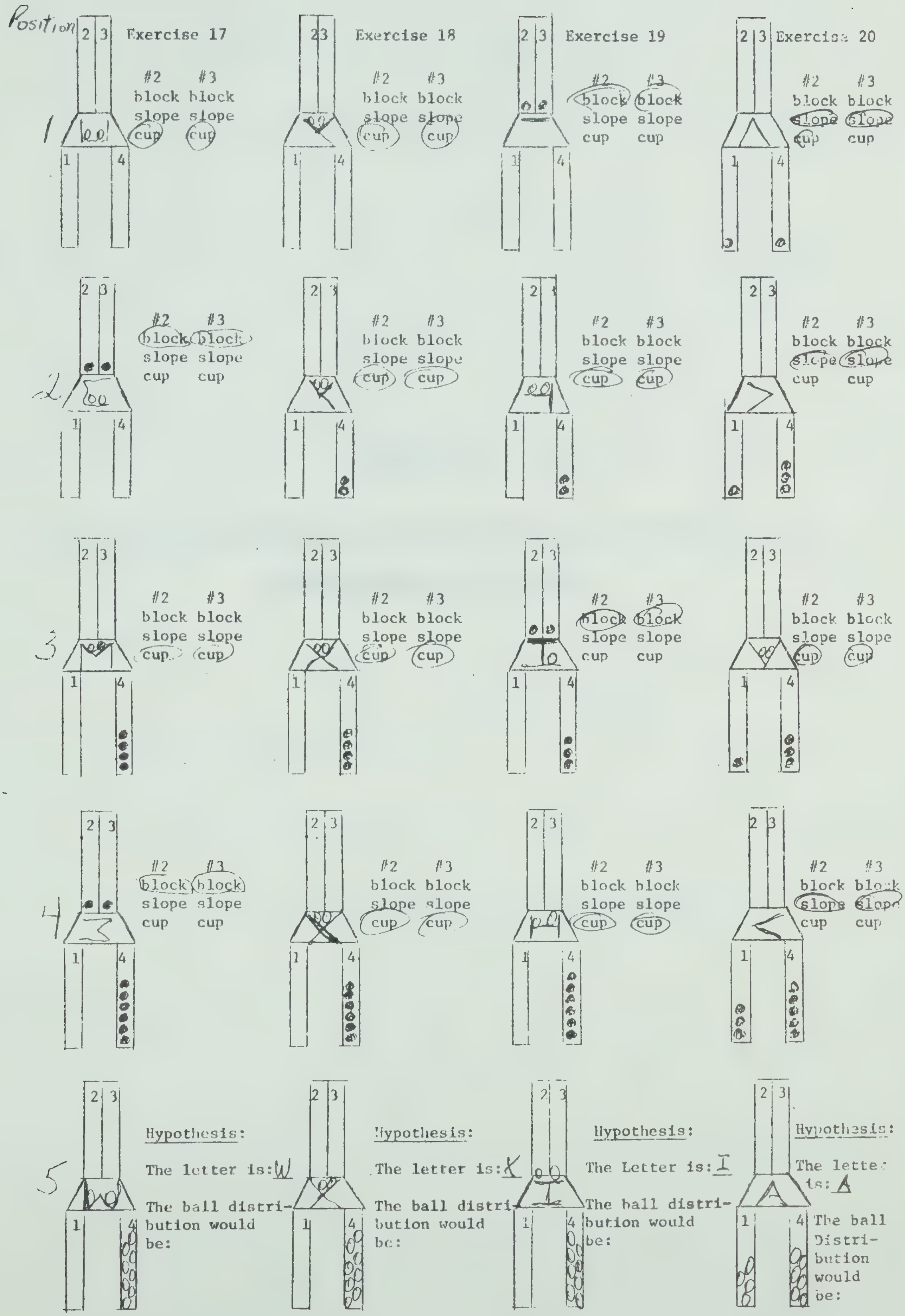
NAME______________________________ SCHOOL______________________________ AGE________




Position
Exercise 9
Exercise 10
Exercise 11
Exercise 12
1
#2 #3
block block
slope slope
cup cup
2
#2 #3
block block
slope slope
cup cup
3
#2 #3
block block
slope slope
cup cup
4
#2 #3
block block
slope slope
cup cup
5
Hypothesis:
The letter is: H
The ball distribution would be:
Hypothesis:
The letter is: U
The ball distribution would be:
Hypothesis:
The letter is: C
The ball distribution would be:
Hypothesis:
The letter is: M
The ball distribution would be:



Position
Exercise 13
Exercise 14
Exercise 15
Exercise 16
1
2
3
4
5
#2 #3
block block
slope slope
cup cup
Hypothesis:
The letter is: V
The ball distribution would be:
Hypothesis:
The letter is: Z
The ball distribution would be:
Hypothesis
The letter is: J
The ball distribution would be:
Hypothesis:
The letter is: P
The ball distribution would be:


Position
Exercise 17
Exercise 18
Exercise 19
Exercise 20
1
#2 #3
block block
slope slope
cup cup
2
#2 #3
block block
slope slope
cup cup
3
#2 #3
block block
slope slope
cup cup
4
#2 #3
block block
slope slope
cup cup
5
Hypothesis:
The letter is: W
The ball distribution would be:
Hypothesis:
The letter is: K
The ball distribution would be:
Hypothesis:
The Letter is: I
The ball distribution would be:
Hypothesis:
The letter is: A
The ball Distribution would be:

APPENDIX D

VALIDATION QUESTIONNAIRE, *INFERENCE TEST* AND *HYPOTHESIS TEST*

INSTRUCTIONS FOR JUDGING

VALIDATION QUESTIONNAIRE

INFERENCE TEST AND HYPOTHESIS TEST

INSTRUCTIONS FOR JUDGING

You have been asked to participate in a study of the inferring and hypothesizing abilities of junior high school students. As your part in the study, you are asked to record your opinion as to how well the two tests, enclosed with this questionnaire, meet certain standards. You are asked to record your opinion on the attached sheets and return them in the return envelope provided.

1. Please judge the instructions for:

clarity:
1. unclear (need major revision)
2. clear (need minor revision)
3. fine (need no revision)

SUGGESTED REVISION:

comprehensiveness:
1. too little or too much (need major revision)
2. just enough (need minor revision)
3. enough (need no revision)

SUGGESTED REVISION:

2. Please judge the items as a whole as being:

a. too easy (trivial for junior high students)
b. easy (most students were not challenged)
c. about right (some students found it hard but most could answer the test)

d. too difficult (very few students can handle the test)

3. Please judge the test format as being:

   a. interesting

   b. uninteresting

   c. waste of time

4. Please judge each item in terms of how well it meets the following standards and is:

   appropriate for inclusion

   inappropriate and should be changed

Inference Test

To answer this item students should exhibit at least one of:

1. arrive at a conclusion about an observation
2. Identify the important factor in each item that is the size and characteristic shape of the target
3. Relate the observation (the pattern) to a given conclusion (the three basic shapes)
4. Recognize that there may be more than one inference (target shape and size) that explains the given observation

Hypothesis Test

To answer this item students should be able to exhibit at least one of the following behaviours:

1. Group a number of inferences about observations into a general explanation (identify the letter-target)
2. Distinguish between the inference about an observation and the end result, the letter-target.

3. Identify the hypothesis that is the result of considering the related inferences

4. Test the hypothesis by predicting the next observation if the experiment is continued.

JUDGING SHEET

(please return in envelope provided)

INFERENCE TEST

1. Instructions had enough information YES NO (check one)

2. Items as a whole were a) too easy b) easy c) about right

   d) too difficult

3. The test was a) interesting b) uninteresting c) waste of time

4. Please identify those items you feel are inappropriate and should be changed.

1	10	19	28
2	11	20	29
3	12	21	30
4	13	22	31
5	14	23	32
6	15	24	33
7	16	25	34
8	17	26	35
9	18	27	36

Suggested changes:

JUDGING SHEET

(please return in envelope provided)

HYPOTHESIS TEST

1. Instructions have enough information YES NO (check one)

2. Items as a shole are a) too easy b) easy c)about right

   d) too difficult

3. The test was a) interesting b) uninteresting c) waste of time

4. Please identify those items you feel are inappropriate and should be changed.

1	6	11	16
2	7	12	17
3	8	13	18
4	9	14	19
5	10	15	20

Suggested changes:

APPENDIX E

INSTRUCTIONS TO TEACHERS

INSTRUCTIONS TO TEACHERS

The schedule of testing and time requirements

1. SCAT test - takes 70 minutes plus a 5-minute instruction period on using the scoring sheets.

2. General Science Test - takes 35 minutes plus a 5-minute instruction period on using the scoring sheets.

3. Inference Test - takes 20 minutes plus a 10-minute instruction period on the format of the test.

4. Hypothesizing Test - takes 30 minutes plus a 10-minute instruction period on the format of the test.

Comments: The tests are to be administered in the given sequence. Within a week from the administering of the SCAT, the hypothesizing test should be completed. Try not to administer more than two tests on the same day. In other words, the SCAT test, which will take two 35 minute periods or a single 80 minute period should be administered on the first day of testing, the General Science Test could be administered alone or with the Inference Test on the second day in two 35 minute periods or one 80 minute period and the Hypothesizing Test administered alone or with the Inference Test on the third day - a total of 5, 35-40 minute periods or 2 1/2, 80 minute periods for testing.

B30138